Ability Grouping and Student Performance: Experimental Evidence from Middle Schools in Mexico*

Matias Busso Veronica Frisancho

December 14, 2022

Abstract

This article relies on a large-scale field experiment in Mexico to measure the effects of two ability-grouping models (tracking and bimodal/heterogeneous groups) on student learning outcomes during middle school. Both strategies yielded an average learning gain of 0.08 of a standard deviation. We find larger and more persistent effects among initially high-achieving students and no significant effects among low-achievers. Students in top tracking had everything going for it: a concentration of high-performing peers and a very homogeneous classroom that facilitated the teacher's work and increased students' effort levels. Bimodal classes fostered greater effort levels among top students while teachers induced less competition and allocated more time to practice and feedback activities in detriment of lecture time. Our results support the allocation of students to homogeneous classes to maximize performance gains among top students, without hurting low achievers. Fostering inclusive learning among weaker students would require complementary investments under both models.

Keywords: Peer effects, tracking, bimodal classes, middle school, field experiment.

JEL Codes: C93, I21, I28, O15.

*Busso: Research Department, Inter-American Development Bank (mbusso@iadb.org). Frisancho: Research Department, Inter-American Development Bank (vfrisancho@iadb.org). We are very grateful to the Autoridad Educativa Federal de la Ciudad de Mexico of the Secretaria de Educacion Publica de Mexico. In particular, we thank the support throughout the project of Sofialeticia Morales Garza, Maria del Carmen Ayllon, Arturo Velatti, Norma Alegria, and Homar Ortiz. Julian Martinez Correa, Juanita Camacho, Jonathan Karver, Nelson Oviedo, and Dario Salcedo provided excellent research assistance at different stages of the project. We also thank Janet Jimenez for locally managing the project. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the the Inter-American Development Bank, its Board of Directors, or the countries they represent.

1 Introduction

Prescriptions to improve learning at the primary and secondary school levels vary substantially in terms of the inputs targeted and their relative cost effectiveness (Evans & Popova 2015). Unfortunately, budget constraints often limit the amount of resources that governments can channel to schools. A cost-effective way to improve learning entails the reallocation of students across classes to exploit positive peer effects.

While the literature on peer effects is quite extensive, considerably less work has been done to study the impact of systematic sorting of students into classes based on academic performance. This paper measures the impact of two prominent classroom-grouping strategies: tracking (i.e., sorting students by initial performance) and bimodal/heterogeneous classes (i.e., grouping weak and strong students together in the same classroom). We designed and implemented a large-scale randomized controlled trial involving a representative sample of 171 public schools in Mexico with almost 40,000 students and more than 500 teachers. While some school systems in the United States, have been tracking students at different schooling levels for decades, there is limited evidence on the impact of this allocation strategy that relies on credible exogenous variation and a large and representative sample (Duflo et al. 2011). Similarly, despite the recent popularity of pedagogical approaches that value diversity and heterogeneity, the evidence available on the impact of heterogeneous classes is limited to the college level (Carrell et al. 2013). Our paper contributes to an important but scarce experimental literature that studies the effects of controlled grouping on academic performance.

We conducted our experiment in a real-life setting and exposed students to substantial changes in the classroom distribution of peers' initial performance. We embedded our experiment in the context of a centralized system that governs admission into public middle schools in Mexico City. We manipulated the allocation of entering students to classes by randomly assigning schools to three treatment arms: a control group that randomly allocated students to classes, and two treatment arms that allocated students under the tracking and bimodal grouping strategies. Classroom composition was determined based on standardized measures of performance at the beginning of middle school, before students actually met. Intensity of exposure to the two treatments was high, as students in the Mexican system spend all school day in the same class with the same group of peers. Exposure was also sustained, as the experimental allocations were not systematically affected between seventh and ninth grade: the three years of middle school. Moreover, our experimental manipulation of classroom composition was not paired with exogenous changes in other inputs from the students' production function of educational achievement. Thus, we are able to isolate the effect of each grouping strategy, absent any complementary change in inputs.

Our paper relies on multiple data sources that allow us to accurately measure changes in academic achievement over time, as well as short-run adjustments of both students' and teachers' behavior. We administered a standardized test and a survey at the end of the first academic year of exposure to the treatment. The standardized exam was produced by the same unit of the Ministry of Education that produces the middle school admission exam on an annual basis. Students' surveys collected data regarding student absenteeism, study habits, risky behavior, classroom dynamic and disruptive behavior, peer academic support, parental support, and existing social networks. A teacher survey was also administered at the end of the first year of exposure to collect information on teaching efficacy, coverage of the curriculum, use of differentiated activities by students' ability levels, time usage in the classroom, and effort. We administered another standardized exam at the end of the second year of exposure to the treatment. We also rely on school administrative records to capture whether students graduated on time from middle school at the end of the third year.

Our results support the existence of average performance gains under both grouping strategies, with clear advantages for high-achieving students. At the end of seventh grade, the first academic year of exposure, both grouping strategies resulted in similar average performance gains of about 0.08 of a standard deviation. The largest learning gains accrued among top-performing students who were placed in tracked classes (0.18 of a standard deviation); top-performing students who were placed in bimodal classrooms (0.13 of a standard deviation); and medium-performing students who were placed in tracked classes (0.06 of a standard deviation). Learning gains were only persistent among top-performing students under both allocation models. At the end of eighth grade, high-performing students in tracked and bimodal classes still exhibited significant performance improvements relative to their counterparts in the control group (0.13 and 0.09 of a standard deviation, respectively). By the end of middle school, both grouping strategies yielded small yet significant effects on the probability of graduating on time from school among higher-performing students. Relative to their counterparts in the control group, low-performing students did not experience any significant performance changes during their middle school trajectories.

Our design and data allow us to study teachers' and students' endogenous responses to the change in classroom composition. In our experiment, schools were informed about their participation in a pilot program but we did not disclose the treatment assignment to school staff, teachers, nor students. Relying on survey responses at the end of seventh grade, we find that teachers' leveraged the comparative advantage of each group-formation strategy. Regardless of the initial performance level of the class, teachers in tracking schools facing more homogeneous classrooms, increased their lecture time and were able to cover significantly more of the curriculum relative to control schools. This effect was stronger in medium- and high-performing classes. Similarly, teachers dealing with bimodal classes seized the advantages of heterogeneous groups by maximizing the opportunity for students to interact by increasing the time allocated to practice and feedback and reducing the incentives to compete among peers.

Survey records reveal that both treatments induced important yet differential changes on students' effort and disruptive behavior, while there was no impact on their risky behavior. First, bimodal classes showed significant improvements in student effort and a reduced number of absences, particularly among high achievers. Average effort was unaffected in tracked groups, but students in high-performing classrooms did exert more classroom effort. Second, low-performing tracked groups seemed to have been more difficult to manage as they record an increase in the level of disruptions. In turn, bimodal classes spent significantly less time dealing with students' disruptions. Third, we fail to identify any significant impact of either group-allocation model on risky behaviors such as violence, extreme disruptive behavior, or substance abuse (i.e., smoking or drinking alcohol).

Low-achieving students who were in tracked classrooms experienced no positive impact on performance and evidenced no increase in effort —findings that align with the evidence that these students were subject to increased exposure to disruptions. However, it is puzzling that low achievers in bimodal groups did not learn more relative to the control group since they had several advantages in their favor: the right endogenous responses from teachers and students were triggered while plausible channels negatively affecting low achievers (i.e., increased risky behaviors) were ruled out. Analysis of students' self-reported friendship networks suggests a plausible explanation. Students in bimodal classes differentially sorted into friendships. On average, low-performing students have a modest tendency to befriend fellow low achievers rather than high achievers. Thus, the higher efforts levels exerted by top performers in bimodal classes may not generate enough spillovers to produce performance gains among low-achieving students.

Our results suggest that tracking maximizes performance gains among top students without hurting low achievers. Complementary policies such as teacher training to better support weaker students as well as remediation programs, delivered during the school day (e.g., Alvarez-Marinelli et al. (2021)) or after school (e.g., Muralidharan et al. (2019)), may be required to foster performance gains among low achievers in bimodal and low-tracked classes. An advantage of tracking is that is facilitates the targeting of complementary investments.

This paper contributes to a large literature that studies peer effects in the educational setting. Identification of peer effects is complicated by two main issues: selection or endogenous group membership and simultaneity (the so-called reflection problem, Manski (1993)).

In recent years, considerable progress has been made to overcome these issues. Some have sought to leverage natural experiments involving variation in the allocation of students to groups (Carrell et al. 2009, Lavy & Schlosser 2011, Imberman et al. 2012). Others have tried to identify peer effects through social networks (Bramoullé et al. 2009, De Giorgi et al. 2010). These efforts have contributed to enhancing understanding of specific channels underlying peer effects. They are, however, limited in their ability to estimate how different group-allocation strategies will affect learning and behavior when changes in peer composition are introduced at a large scale. Natural experiments or network overlaps exploit marginal variations in peer composition that are not likely to lead to teachers' or students' endogenous responses that are equivalent to those triggered by larger-scale ability-grouping strategies.

More specifically, our paper contributes to the somewhat scant experimental literature on controlled classroom grouping. The practice of grouping students of similar achievement levels into classes (i.e., through tracking or streaming) is a long-standing tradition in education (Turney 1931, Martin 1927).[1] However, rigorous evidence on the effectiveness of tracking on learning has been elusive.[2] Earlier work relied on small-sized randomized controlled trials in developed countries and yielded null impacts on performance (Slavin 1987, 1990, Betts & Shkolnik 2000). In the context of a primary school class-size-reduction experiment, Duflo et al. (2011) studied the impact of tracking in elementary schools in Kenya and found performance gains both among low- and high-achieving students. The literature on the effects of heterogeneous classes is much thinner, even though variation in peers' ability naturally occurs in many settings (e.g., rural multi-grade schools or Montessori schools). Using observational data, Carrell et al. (2013) found that the group allocation that maximizes performance gains among low performers is one that mixes them with top performers. However, when Carrell et al. (2013) set out to experimentally evaluate performance gains under this model, they found a negative and significant treatment effect for the students they intended to help, exposing the limitations of natural experiments to predict the effects of controlled grouping strategies. Booij et al. (2017) undertook an alternative approach: they introduced large and exogenous variation in peers' performance distribution across groups, estimated flexible reduced-form models of peer effects, and generated performance predictions under different allocation strategies.

Our paper builds on these studies and brings several advantages over previous experi-

[1] In this paper, we refer to ability-grouping strategies as those that are implemented in the classroom within a school and are kept constant for all subjects throughout the school year. Slavin (1987) discusses a typology of other ability-grouping strategies in education such as tracking students to different schools (e.g., vocational or academic, normal or gifted, etc.), creating special classes for low achievers (e.g., remedial education), and using within-class ability grouping, among others.

[2] Betts (2011) reviews the earlier theoretical and empirical literature assessing the effect of tracking on students' outcomes.

mental work. First, our design allows us to isolate the effect of controlled grouping, without affecting any other input such as class size or teachers' characteristics. Second, we rely on standardized measures of performance captured before students meet to allocate them to classes. Because these measures are observed when students apply to middle school, they are not distorted by previous interactions with school peers. Third, we rely on rich data sources that allow us to measure performance improvements using a standardized instrument, and to examine behavioral changes among teachers and students. Fourth, we implement ability-grouping models for the education level at which peers become the most influential: middle school.

Our paper is also relevant in terms of its implications for scalability. Except for Duflo et al. (2011), other experimental studies tend to rely on data from only one school. Our experimental sample is not only large, but is also representative of public schools in Mexico City. We partnered with Mexico City's government and made sure that the implementation efforts were undertaken by the regular, institutional actors, with limited guidance from our end. We also refrained from introducing additional inputs in the performance production function. The ultimate goal was to help devising a light-touch system that could easily be implemented and maintained in school districts without the need for additional resources.

The remainder of this paper proceeds as follows: Section 2 describes the experimental design, data sources, and assesses the internal validity of our study. Section 3 presents the empirical strategy and main experimental results on academic performance. Section 4 shows evidence on the effects on teachers' and students' behavior. Section 5 offers conclusions and discusses policy implications.

2 Experimental Design

2.1 Setting

The Mexican public school system offers four levels of schooling: preschool (from age three to kindergarten), elementary school (grades one through six), middle school (grades seven through nine) and high school (grades ten through twelve). Our experiment targeted entering cohorts to public middle schools in Mexico City, one of the largest school districts in the country, serving more than two million students. Approximately 160,000 students enter middle school each year and about 85% of them attend one of the 832 public middle schools in Mexico City.[3]

[3] The number of schools provided corresponds to combinations of school buildings and shifts, since one building may host two different shifts. Middle school education in Mexico City is provided through two types of facilities: academic schools, which represent 72% of middle schools, and technical track schools,

A salient feature of Mexico City is that students are assigned to their middle schools via a centralized allocation mechanism which is implemented in six steps: (i) between January and February of each year, elementary school students submit a ranked ordered list of up to three schools; (ii) in early June, students take a standardized admission test (Section 2.3 provides more details on this instrument); (iii) at the end of June, students are allocated to schools following the Boston mechanism (Calsamiglia & Güell 2018). In an initial round, all students compete for their first choice based on their score on the standardized test.[4] Those who are not assigned, move onto the next round and compete for their second choice. The process is repeated a third time for students who remain unmatched. At the end of the allocation process, unplaced students are assigned to a school with available seats and that is located near the candidate's top school in her submitted ranked ordered list; (iv) at the beginning of their summer break, students are informed about their school assignment; (v) in mid-July, students who are dissatisfied with their school assignment may request a change. Requesting a change requires the student to forego the seat they had been originally assigned and to search for an available seat. In the year prior to our experiment, approximately 18% of the applicants from our experimental sample requested a change and 7.5% were granted a seat in a different school; (vi) final allocation results are announced early in August. The allocation mechanism induces certain degree of positive assortative matching of students to middle schools (based on the admission test scores), but sorting is limited by the cost of commuting across a very large and densely populated city. Indeed, the ratio of the between-school to the within-school variances in the standardized admission score is about 1.1.

Once all applicants are placed, the roster of incoming students is sent to school principals. Principals allocate students to groups, typically without following any preset protocol. According to self-reported pre-treatment data collected among principals, fewer than a third took into account information from students' past performance (i.e., primary school GPA) to allocate students to groups, and none of them relied on the admission standardized test scores. In general, principals' main goal when forming groups was to maintain some balance, both in terms of the age distribution and the sex composition.

Throughout the school year, students in a group spend the full school day in the same classroom: teachers rotate across classes and students receive all subject lessons together. While group composition can be affected by the entry and exit of students and/or parents'/teachers' requests to make specific group changes, it is not unusual to keep the same

which represent the other 28%. Both cover the same curriculum, but technical schools also offer vocational training. In academic schools, each shift is managed by different principals; in technical schools the same principal manages both shifts. In our sample, only 19% of the schools offer both daytime and evening shifts.

[4] Ties are broken by giving preference to those applicants who have a sibling enrolled in the school at the time of the application and to those who live closer to the preferred school.

classmates during the three years of middle school.

2.2 Research Design

Intervention. Our intervention affected the way students were assigned to classrooms. Once the centralized allocation system had assigned students to schools, we received the final placement results, each student's admission test scores, and the number of seventh grade groups/classrooms in each school. Based on these data, we defined three different models of group formation:

(a) Tracking: Students were sorted according to their admission test score and grouped into classrooms based on their relative standing at the school level and on class sizes. For example, if 90 students are assigned to a school with three classrooms, the 30 lowest-achieving students were assigned to group A, the 30 medium-achieving students to group B, and the 30 highest-achieving students to group C. Whenever ties emerged, we randomized the allocation of students at the score cutoff to adjacent groups.

(b) Bimodal: Within each school, students were first divided into terciles based on their admission test score. Then, two types of classrooms were formed: bimodal and homogeneous classrooms. Bimodal classrooms included students from the low- and high-achieving terciles. Students in the medium-achieving tercile were assigned to homogeneous classrooms. Using the same example above, with 90 students and three classrooms, we assigned 15 of the 30 lowest-performing students and 15 of 30 highest-performing students to group A; the 15 remaining lowest-scoring and 15 remaining highest-scoring students were assigned to group B; all medium-performing students were assigned to group C.[5]

(c) Control: Students in control schools were allocated to classrooms at random.[6]

Note that the algorithm implemented to allocate students under each treatment arm reproduced the sex composition in the school at the classroom level.

Randomization. We defined the experimental sample of schools by imposing five restrictions on the universe of public middle schools in Mexico City. First, we dropped the very

[5] With larger student bodies and more classrooms comes greater variance of the entry score. To ensure that all bimodal classrooms were comparable in terms of the initial distribution of peers, we subdivided each tercile into three subgroups and randomized an equal number of students from each of these subgroups into all bimodal classes. Homogeneous classrooms in bimodal schools are similar to medium-achieving-level classrooms in tracking schools and, for that reason, we leave them out of the analysis.

[6] In Busso & Frisancho (2021) we exploit this random assignment to analyze the gendered effects that the presence of high-achieving peers in the classroom has on girls' high school placement outcomes.

best performing schools (those in the top 10% of the average admission test score distribution in the year prior to the experiment). Second, we left out schools that seated fewer than 80 students in seventh grade. Third, we dropped schools with low dispersion in the admission test scores. That is, we dropped schools with a coefficient of variation in the admission test scores of under two-thirds. Fourth, we excluded those schools whose entering cohorts had a relatively high share of students who were older than 15 and/or who had special needs. Finally, in the case of technical schools with more than one shift, we exclude one at random to minimize the risk of noncompliance with treatment; these schools share the same principal across shifts and different treatment assignments across shifts would have raised contamination concerns. Thus, the final eligible universe included 452 middle schools from which we selected 171 schools at random.

We stratified the eligible universe of schools by school type (general or technical), quartiles of the average performance in a previous national exit exam[7] and shift (morning or afternoon). Within each of these cells, we ranked schools by the total number of enrolled students, and we formed strata of size three. We then randomly assigned each school within each stratum to the control, tracking, or bimodal treatment arms.

We conducted a single-blinded experiment. Principals knew they were participating in a study, but they did not know their treatment status. Teacher allocation was uncorrelated to group formation in the school. Teachers had been assigned to classrooms by the principal at the end of the previous academic year, before classroom composition was released. We implemented the three allocation models depending on the results of the randomization of schools into treatment arms. For all schools in our sample, we produced a list of entering students with their corresponding group assignment. The Ministry of Education handled communication with schools' principals to ensure that the group assignment was implemented following the lists we had produced.[8]

Timeline. Figure 1 shows a timeline of school-related activities (in italics) and evaluation-related activities (in bold) between 2015 and 2018. We generate group assignments among the cohort entering seventh grade right before the beginning of the 2015 academic year. Classes started at the end of August 2015 and finished in May of the following year. We collected data on students' outcomes relying on survey instruments at the end of grade seven. At the same time, we surveyed teachers using an instrument that tried to measure teaching strategies and self-perceptions about effort and efficacy.

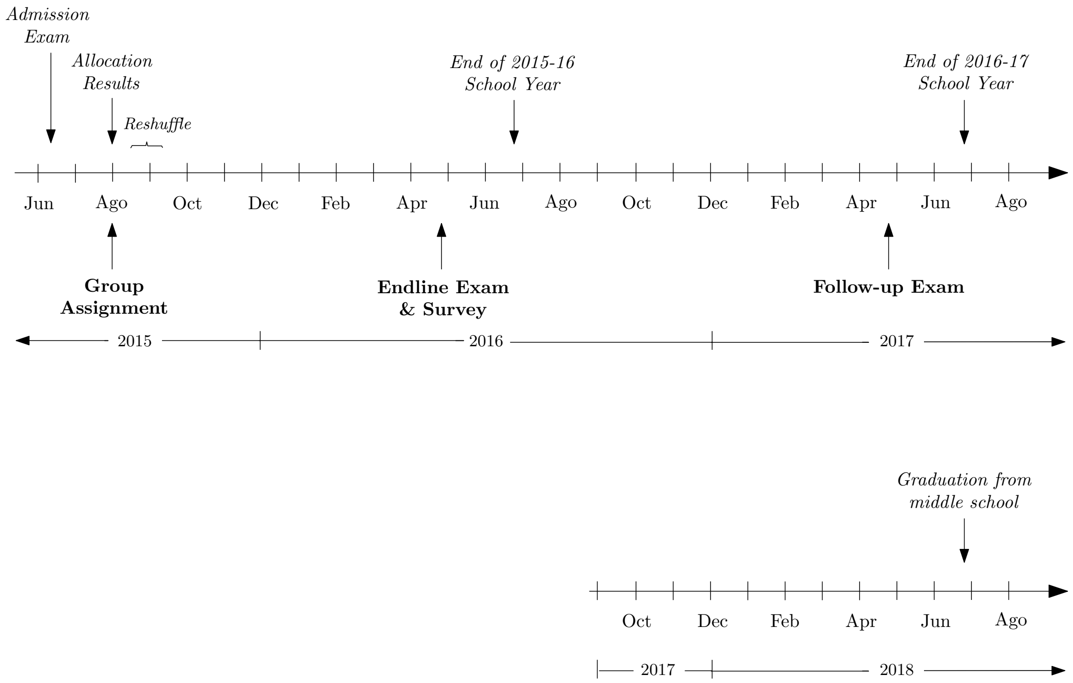
We measured academic performance at the end of grades seven and eight using an in-

[7] In 2013, Mexico administered a standardized test (School National Achievement Test, called ENLACE for its Spanish acronym) to all middle schools in the country.

[8] In extreme situations when special accommodations were required for specific students, principals had the option to inform the Ministry and deviate from the original classroom assignment.

strument that mimicked the middle school admission test. We also obtained access to administrative records corresponding to students' full trajectories throughout middle school, between grades seven and nine. In particular, we observe whether students graduated from middle school on time – that is, by the end of the 2017-2018 school year.

Figure 1: Study Timeline



2.3 Data

Administrative records. We use students' application records, which are assembled as part of the centralized admission process for public middle schools. These data, which serve as a baseline, feature the admission test score, elementary school cumulative GPA, and basic socio-demographic information about the student (including gender, age, number of siblings already enrolled in the preferred schools, and whether the student has any type of learning or physical disability). The registration form also includes household socioeconomic characteristics such as parents' ages and education levels, number of household members, and whether (or not) the student lives with both parents. A second source of administrative records contain data on students' progress through middle school; these allow us to assess whether students graduate on time. These records also help us to identify student turnover and group changes throughout the middle schools in our sample as well as student absenteeism as recorded by schools.

During the randomization stage, we relied on school-level data which include the number of students enrolled and average academic performance in previous school years, as measured by school results in a centralized national exit exam administered in 2013. All these records come from the Mexican school census and the Ministry of Education.

Standardized test scores. In 1989, the federal government decided to develop a new standardized test (IDANIS for its Spanish acronym) to measure student preparedness for middle school. The test, administered at the end of the last year of elementary school (Grade 6), tests students on three domains and five sub-domains: literacy (reading comprehension and writing) mathematics (arithmetic, geometry) and abstract reasoning. The test consists of 60 multiple-choice questions with varying degrees of difficulty, worth one point each. There is no negative marking. In 1996, Mexico City started administering the test to the entire population of elementary school students; since then it has been used as the main ranking criterion in the centralized school allocation mechanism discussed in Section 2. Therefore, at the time of the experiment, there was a vast repository of testing materials that had been properly piloted and vetted for quality.

The standardized tests administered in the follow-up rounds of this study were produced by the same unit of the Ministry of Education that produces the annual admission exam. The members of that unit developed exams comparable to the admission test to measure students' achievements at the end of grades seven and eight on the same domains and sub-domains included in the admission test. The tests are psychometrically valid. All the Cronbach's alpha coefficients for each domain are higher than 0.7, and their estimated difficulty and discrimination parameters are statistically significant.[9]

Surveys. We collected students' surveys at the end of grade seven. These surveys gathered information on student behavior that can allow us to better understand how peer effects operate within the classroom. Using scales previously defined and validated in the literature, we collected self-reported data regarding student absenteeism, study habits, risky behavior, classroom dynamic and disruptive behavior, peer academic support, parental support, and existing social networks.[10] To measure social networks, we asked students to list the first and last names of their three best friends in the classroom.

At the end of grade seven, we also surveyed teachers in an attempt to measure teaching practices and their self-perceptions about effort and efficacy levels. We included scales that

[9] The instrument design does not include a set of items that would allow one to anchor the test from one year to the next. Thus, tests are not strictly comparable over time. Section 2 of Appendix A (Data) provides the psychometric assessment of each test item (estimated difficulty and discrimination parameters) relying on a dichotomous item response theory (IRT) two-parameter model. We estimate these parameters for the treatment and control students and find virtually identical properties for all items in both samples.

[10] The scale used to measure study habits comes from the Student Engagement in Schools Questionnaire (Hart et al. 2011). Scales to measure disruptive behavior are extracted from Patterns of Adaptive Learning Scales (Midgley et al. 2000). The peer academic support scale comes from the Classroom Life Instrument (Johnson et al. 1983).

allowed us to measure mastery and performance approaches and teaching efficacy (Hart et al. 2011), coverage of the curriculum, use of differentiated activities by students' ability level, time usage in the classroom, and effort.[11,12]

Sample. Our experimental sample consists of 171 schools, with 907 classrooms in total. Administrative records are available for all students in the schools that participated in the experiment ($N = 32,324$). At the end of grades seven and eight, standardized tests and surveys were administered (for costs reasons) to students in a subset of classrooms: all classrooms in tracking schools, three classrooms chosen at random from each control school, and three bimodal classrooms chosen at random from each bimodal school. We did not collect survey or administered tests in homogeneous classes from the bimodal treatment arm as they are similar to medium tracking classrooms. Consequently, all effect sizes reported for bimodal schools actually correspond to bimodal classes (i.e., those truly heterogeneous) and do not include the effects on medium-performing students.

The total survey samples in the first and second follow-up rounds correspond to 19,762 and 16,778 students, respectively. The teachers' survey was applied to all math and Spanish teachers who were delivering instruction in these courses during the first academic academic year of the experiment.[13]

2.4 Experimental Validity

Balance. Students in tracking, bimodal and control schools did not, on average, differ in terms of any pre-treatment characteristic. Figure 2 presents sample means of students', schools' and teachers' characteristics in tracking, bimodal, and control schools. It also plots the distribution of p-values of a test of equality of sample means between tracking and control (solid markers) and between bimodal and control (hollow markers) schools estimated using an ordinary least squares model that controls for strata fixed effects and clusters standard errors at the school level. Students in all treatment arms have similar initial performance levels as measured both by the standardized admission test score and their elementary school GPA. Schools across different arms are also very similar in terms of performance (measured by a national standardized test) and performance heterogeneity. We also rule out differences in terms of number of classrooms and average class size: the average school has about 5 classrooms with 34 students each. Teachers are also similar across treatment arms. The

[11] The questions on teachers' time usage were based on the Teaching and Learning International Survey by the Organisation for Economic Co-operation and Development (OECD) (http://www.oecd.org/education/talis/). All students' and teachers' scales used in this paper are described in more detail in Section 3 of Appendix A (Data).

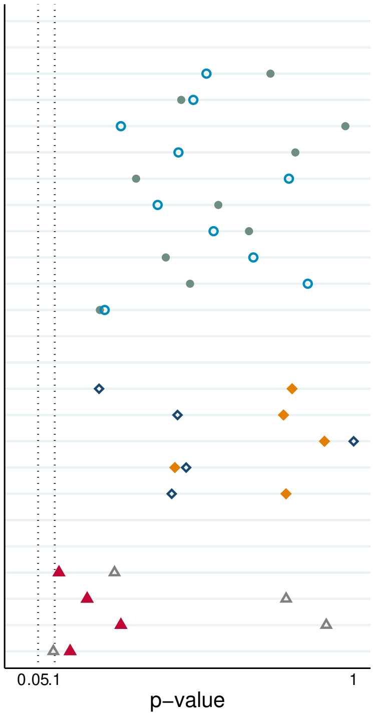
[12] Additional details on the data sources are available in Section 1 of Appendix A. (Data).

[13] See Section 4 of Appendix A (Data) for further details on the number of observations by data source.

only statistically significant difference is that teachers in bimodal schools seem to be slightly less experienced than teachers in control schools.

Figure 2: Pre-Treatment Average Characteristics

	Tracking	Bimodal	Control	Obs.
Panel A. Student characteristics				
Primary GPA	8.451	8.446	8.437	38130
Secondary or higher, father	0.654	0.652	0.644	37873
Extemporaneous, initial	0.045	0.040	0.045	39529
Lives with both parents	0.649	0.640	0.646	37873
Accepted transfer	0.025	0.029	0.028	39529
Male	0.513	0.526	0.519	39529
Secondary or higher, mother	0.746	0.759	0.749	38070
Special needs (USAER or self−reported)	0.024	0.027	0.026	37873
Over 12 years of age	0.076	0.073	0.073	38140
Initial test score	25.505	25.348	25.924	37812
Panel B. School characteristics				
SD initial test score	10.373	10.086	10.443	171
Change in # of students, 2012−2015	0.103	0.050	0.086	171
Number of classrooms	5.316	5.298	5.298	171
Avg. test score (ENLACE) 2013	491.673	484.534	487.239	170
Class size	35.030	34.791	34.400	907
Panel C. Teacher characteristics				
Age	40.784	41.370	42.473	514
Female	0.600	0.630	0.629	514
Level of studies (bachelor's degree)	0.876	0.840	0.844	514
Years of experience	13.481	13.259	14.778	514

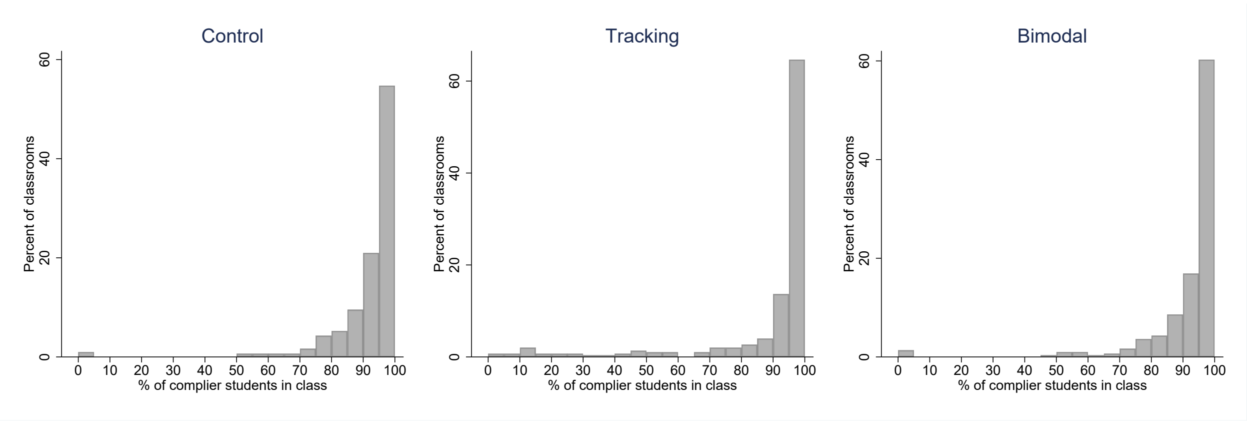


Note: Each panel presents the pre-treatment mean of students' (Panel A), schools' (Panel B), and teachers' (Panel C) characteristics in tracking, bimodal and control schools. The figure on the right plots the p-values associated to the null hypothesis that the mean difference for a given characteristic when comparing tracking (solid markers) / bimodal (hollow markers) to control schools is equal to zero. These tests are based on an ordinary least squares regression using each baseline characteristic as the dependent variable. All regressions include strata fixed effects. Standard errors are clustered at the school level.

Compliance. We worked closely with government officials and principals to ensure that the classroom assignments were implemented following our guidelines. Compliance was high and very similar across all three treatment arms. For each treatment arm, Figure 3 presents the distribution of compliance rates at the classroom level. Among the 907 classrooms in the experimental sample, and irrespective of the treatment arm, seven out of nine classes had at least 90 percent of complying students —and eight out of nine classes had at least 80 percent. That is, the vast majority of students sat in the group we allocated them to.[14] Deviations from perfect compliance can respond to a variety of reasons that are likely to be unrelated to treatment assignment. Field reports record noncompliance due to changes in the student body during the school year or special situations (such as students with physical disabilities who could not go to a classroom in buildings with a second floor).

[14] We cannot reject the null hypothesis that the proportion of classrooms with at least 90 percent of compliance was equal across tracking, bimodal, and control classes. See Section 1 of Appendix B (Experiment)

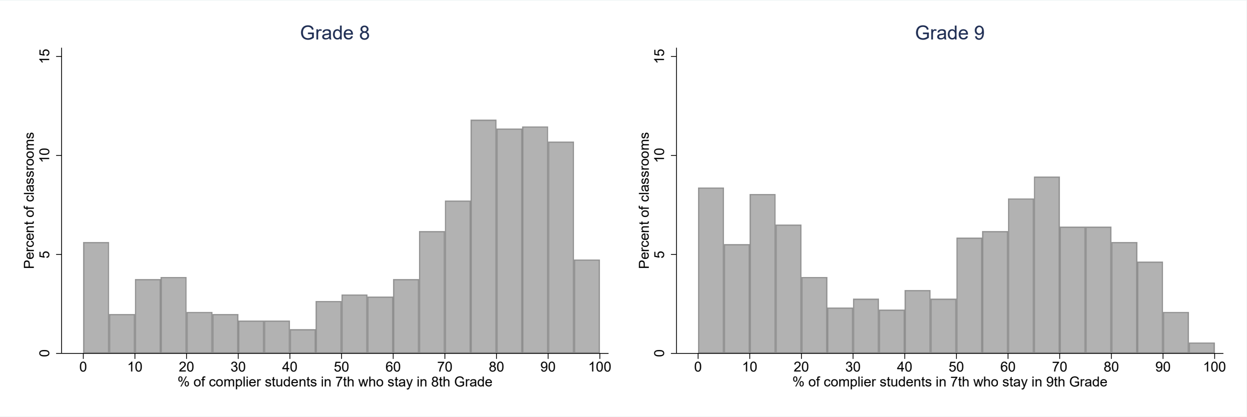
Figure 3: Compliance Levels by Treatment Arm



Note: For each treatment arm, the figures plot the histogram of the percent of students in each classroom who were observed in the administrative records in the classrooms to which they were assigned by the randomization protocol.

Classroom Stability. After initially allocating students to classrooms, the Ministry of Education gave no further instructions related to preserving the groups in eighth and ninth grades. Despite the lack of guidance, classes were not completely dismantled, but they were affected by the normal churn of students, irrespective of the treatment assignment.[15] Toward the beginning of eighth grade, the average share of students in classrooms as originally indicated by our intervention dropped to 63 percent, and by ninth grade to 46 percent.

Figure 4: Group Stability Levels for Eighth and Ninth Grades
(Percent of Classrooms by Share of Treatment Compliers)



Note: The left panel shows the histogram of the percentage of seventh-grade students in each classroom who were still in the assigned classroom by eighth grade. The right panel shows the histogram of percentage of seventh-grade students in each classroom who were still in the assigned classroom by ninth grade.

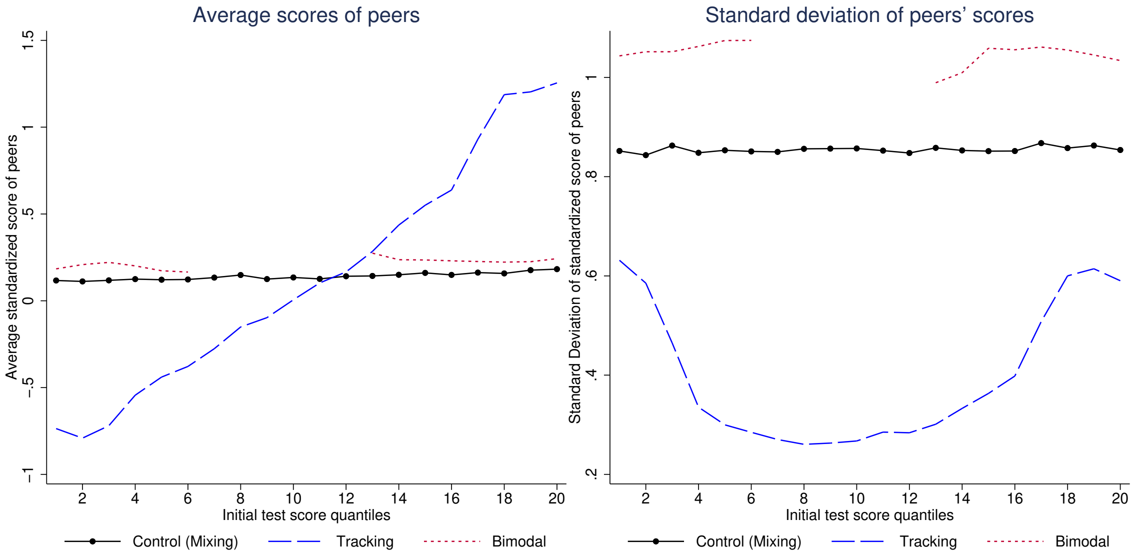
for more details on compliance.

[15] Section 2 of Appendix B (Experiment) shows that the churn of students in eighth and ninth grades is not related to treatment assignment.

Figure 4 shows the histogram of the proportion of students who were still seated according to their initial group assignment throughout grades eight and nine. The histograms reflect that, even though the initial group allocation changed over time, most of the classrooms in grade eight remained approximately the same. While churn is more noticeable by grade nine, the figure shows that there is still an important share of the classrooms that keeps most of the original formation. This is a common practice in Mexican middle schools and is not related to any specific instructions under the experiment.[16]

Changes in Peer Composition. Each treatment arm had a differential impact on the distribution of peers' academic achievement. Figure 5 shows the observed mean (left panel) and the standard deviation (right panel) of peers' admission test scores for students in different ventiles of the score distribution. Since schools vary in terms of the distribution of admission test scores, we standardize scores within school before computing the first two moments of peers' initial academic achievement.

Figure 5: Experimental Variation in Peers' Academic Achievement Distribution



Note: The left figure plots, for each treatment arm, the average standardized pre-treatment test scores of classroom-peers for each student i (by ventiles of the pre-treatment score of student i). The right figure plots the standard deviation of the standardized initial score of classroom-peers of student i (by ventiles of the pre-treatment test score of i).

The average credentials of peers across control classes is consistent with the randomization of students into classes under this arm: an individual student's initial performance is uncorrelated with average group's performance. The standard deviation of the peers' admission tests scores is also uncorrelated with individual test scores in control classes.

[16] Section 3 of Appendix B (Experiment) shows that there is no statistically significant difference in terms of group stability across treatment arms.

Compared to the control group, high- and low-achieving students in bimodal classrooms shared classes with peers of similar average initial achievement. However, they attended school in more heterogeneous classrooms. This was the result of removing middle-performing students from the distribution. In turn, tracking schools induced positive assortative matching in terms of initial scores. Relative to the control group, students in tracking schools attended much more homogeneous classrooms.[17]

The effects of treatment assignment on the distribution of peers in tracking and bimodal schools vis-à-vis those in control schools is still observed in the data by eighth and ninth grades (even if somewhat ameliorated).[18]

3 Experimental Results

We measure the impact of tracking (T) and bimodal (B) classroom grouping strategies on students'/teachers' outcomes by estimating intention-to-treat effects from the following ordinary least squares regression model:

$$Y_{igsq} = \alpha + \beta_T D^T_{sq} + \beta_B D^B_{sq} + \delta X_{igsq} + \omega_q + \epsilon_{igsq} \tag{1}$$

where Y_{igsq} is the outcome of interest (e.g., post-treatment student academic performance) of student/teacher i in group/classroom g in school s and strata q. The impact of each treatment is measured by β_j with $j \in T, B$, the coefficients on each indicator variable denoting tracking (D^T_{sq}) and bimodal (D^B_{sq}) treatment assignment, respectively. All regressions include additional individual and classroom characteristics as controls, denoted by vector X_{igsq}. This vector also includes the baseline value of Y_{igsq}. We also control for a set of indicator variables, ω_q, which identify the stratum q to which school s belongs to. Standard errors are clustered at the school level, the unit of randomization.

Table 1 presents the main results of the controlled grouping intervention. Columns [1] and [2] in Panel A show the treatment effect on the aggregate standardized test score after one academic year of exposure to peers. On average, students in tracking and bimodal schools learned more that students in control schools. Gains among students in tracking were 0.079 of a standard deviation while gains among students in bimodal schools were 0.083 of a standard deviation relative to students in control schools. Both treatment groups

[17] The higher levels of variance for very low and high score ventiles in the tracking group reflect that bottom and top classes were partly populated by "outliers" at both ends of the achievement distribution. The bottom classes in tracking schools host students who did not take the entry exam (i.e., students repeating the grade or who did not register on time to participate in the centralized allocation mechanism) and contribute with imputed zeros. The top classes host the few students who obtained perfect scores.

[18] See Section 3 of Appendix B (Experiment).

experienced gains in math and literacy.

Table 1: Overall Effects of Tracking and Bimodal Classroom Assignment

				By Initial Achievement				
				Low		Medium	High	
Baseline variable	Tracking	Bimodal	Obs.	Tracking	Bimodal	Tracking	Tracking	Bimodal
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
Panel A: 7th Grade								
Total	0.079	0.083	18,795	–0.019	–0.020	0.061	0.185	0.127
	[0.030]***	[0.035]**		[0.041]	[0.041]	[0.035]*	[0.046]***	[0.050]**
Math	0.081	0.089	18,795	–0.005	–0.018	0.027	0.206	0.137
	[0.024]***	[0.026]***		[0.034]	[0.037]	[0.028]	[0.040]***	[0.041]***
Literacy	0.059	0.066	18,795	–0.032	–0.018	0.082	0.124	0.112
	[0.031]*	[0.036]*		[0.044]	[0.042]	[0.039]**	[0.042]***	[0.046]**
Panel B: 8th Grade								
Total	0.041	0.066	13,825	–0.027	0.016	0.007	0.134	0.092
	[0.029]	[0.034]*		[0.045]	[0.045]	[0.032]	[0.041]***	[0.044]**
Panel C: 9th Grade								
Graduation Rate	0.003	0.003	22,612	–0.002	–0.002	0.007	0.004	0.007
	[0.002]	[0.002]		[0.004]	[0.004]	[0.003]**	[0.002]*	[0.002]***

Note: Columns [1] and [2] report estimates of tracking and bimodal treatment effects using equation (1) on each row outcome. Columns [4]-[8] report estimates on tracking and bimodal treatment effects by splitting the sample according to the student's initial achievement (low/medium/high refers to tercile 1/2/3 of the pre-treatment test score school distribution). We did not collect information for students in the medium tercile in bimodal schools. Panel A shows estimates of scores on standardized tests (literacy, math and total/aggregate test score) taken at the end of seventh grade. Panel B shows estimates of scores on standardized tests taken at the end of eighth grade. Panel C shows estimates on graduation on time from ninth grade. Estimators include strata fixed effects. All estimates correspond to equations that control for initial test score, an indicator if the exam score is equal to zero, having the initial score imputed, age, gender, an indicator of whether the student has pre-registration, administrative unit, number of classrooms per school and number of students by group. The total number of observations is 5,973 for the lowest performance tercile, 5,090 for the middle tercile, and 7,732 for the highest performance tercile. Statistically significant at * 10%, ** 5% or *** 1% level. Standard errors clustered at the school level.

Columns [4]-[8] in Panel A present estimates of treatment effects for students in different terciles of the school distribution of the standardized admission test score. We exclude from the analysis the second/medium tercile in bimodal schools because these students were grouped in homogeneous classes within bimodal schools.

The results show that neither tracking nor the bimodal allocation of students to classrooms improved individual performance of students in the lowest tercile of the initial performance distribution. Medium- and high-performing students, on the other hand, experienced a significant learning boost in tracking and bimodal schools. In particular, top students in the third tercile accrued the largest performance gains at the end of seventh grade under both allocation models. In both cases there were performance gains across math and literacy. Relative to the control group, however, learning among the highest-performing students improved the most under tracking. At the end of seventh grade, students in the top tercile who attended tracking classes experience performance improvements equivalent to 0.185 of a standard deviation; by contrast, their counterparts in bimodal classes accrue gains in performance of 0.127 of a standard deviation.

Panel B presents outcomes for eighth grade. Though somewhat persistent, the effects of both tracking and bimodal classes seem to be attenuated as time goes by. At the end of eighth grade, after two academic years of being exposed to tracking and bimodal groups, the average treatment effect estimates remain positive, but statistically significant only among students in bimodal schools. Columns [7] and [8] in Panel B show that top students were the only ones who are able to sustain most of the performance gains accrued during seventh grade. At the end of eight grade, the highest-performing students in tracking classes performed 0.134 of a standard deviation above their counterparts in the control group; students in bimodal classes performed 0.092 of a standard deviation above their counterparts in the control group.

Panel C presents the estimated treatment effects on the probability to graduate on time from middle school. On average, neither tracking nor bimodal classrooms had a significant impact on graduation. However, consistent with the performance results previously discussed, both allocation models yielded small but significant effects on the probability that medium- and high-performing students graduate on time. Relative to their counterparts in the control classes, students from the second tercile under the tracking model experienced a 0.7 percentage point increase in the probability of graduating on time. Similarly, treated students in the third tercile showed improvements in graduation rates equivalent to 0.4 percentage point under the tracking model and a 0.7 percentage point under the bimodal allocation.

Notice that our estimated treatment effects capture the impact of tracking and bimodal classrooms on students' performance *absent* any complementary investments. Tracking is often recommended under the premise that schools could then channel additional inputs and/or existing resources more appropriately to each group. Principals could assign more experienced teachers to low-performing classrooms, or they could target remediation programs to them. Similarly, teachers in bimodal classrooms could be trained to better target instruction in a heterogeneous group and to use cooperative pedagogical tools. None of these complementary investments took place in the context of our experiment. Our design isolates the effects that occur from changing the classroom composition only – allowing for endogenous responses of both students and teachers.

Our experiment was designed building on performance-based grouping strategies that had been previously tested using randomized control trials; albeit in different schooling levels. In particular, Duflo et al. (2011) studied the tracking model in elementary schools in Kenya while Carrell et al. (2013) analyzed the *bimodal* allocation in a U.S. college.

Our estimated average treatment effect for tracking is less than half of that obtained by Duflo et al. (2011) in the case of Kenyan first graders (0.182 of a standard deviation). This difference is driven primarily by the differential impact on low-achieving students: while we found no gains for this group, Duflo et al. (2011) find a positive effect —equivalent to that observed among high-achieving students. This differential impact of tracking on low-achieving students could be explained by at least two important factors. First, the incentives faced by teachers were different in both settings. In our setting, the vast majority of teachers are hired as public-sector employees through permanent contracts. In the Kenyan setting, the share of permanent teachers is around 50%, and effort responses significantly differed across civil servants (i.e., hired under fixed contracts) and contract teachers. In fact, the authors find differential treatment impacts on students' performance depending on the type of contract given to the teacher; low-achieving students of permanent teachers do not seem to benefit that much from tracking, as is also the case in our study. The average gains recorded among low-scoring tracking classes were mostly driven by those assigned to contract teachers. Second, both interventions took place at different grade levels. It is possible that first-grade students react differently to being tracked into a lower-achieving classroom when compared to middle school students. Even though in both cases students were not informed that they had been assigned to a low, tracked classroom, students' behavior could have endogenously changed in different ways across settings. For instance, a higher concentration of low-achieving peers in middle school could have led to a greater degree of disruptions during lectures, whereas the scope for disruptive behavior may have been more limited among first-grade students.[19]

Our results for bimodal schools also differ from those found by Carrell et al. (2013) for students in the U.S. Air Force Academy. While the authors find a null average treatment effect, we find average performance gains in our bimodal classes. Their results were driven by a negative and statistically significant effect for low-achieving students, a positive and statistically significant effect for students at the middle level of achievement, and the absence of gains for top students. The differential impacts between our study and those of Carrell et al. (2013) can be partially explained by differences in exposure. Students in bimodal schools in our experiment spent almost all the time with the same peers for the totality of the school year (and in many cases for the totality of middle school). Air Force Academy students were grouped into squadrons that became their main peer reference group in terms of social and academic interactions, but core courses were taught in small sections that mixed students from all squadrons. The relatively higher exposure in our setting allows students and teachers to better adapt to and reap the gains from a bimodal classroom.

[19] Appendix Figure 1 shows the negative correlation between disruptive behavior and pre-treatment test scores of students in the control group.

4 Mechanisms: Behavioral Responses

4.1 Peer Effects in the Classroom

The changes in classroom composition that we introduced affected learning by potentially triggering several endogenous behavioral responses among students and teachers. First, students' learning can be directly affected by the change in the pool of peers. For instance, students may benefit from the externalities generated by a higher concentration of high-achieving peers who are more likely to ask better questions, and are less likely to disrupt the class. Second, students' behavioral responses indirectly affect their peers' learning. On the one hand, students' effort levels may vary depending on how incentives to compete are affected by the change in the classroom composition. Recent studies suggest that relative ranking matters in certain settings. Changes in the mean and variance of the distribution of initial performance can thus affect students' aspirations and, consequently, their effort levels (Tincani 2018, Murphy & Weinhardt 2020, Delaney & Devereux 2022). Relatively stronger students can also become role models and induce greater levels of effort among their peers (Balestra et al. 2021, Cools et al. 2021). On the other hand, changes in the classroom composition may lead students to change the way in which they interact. Previous studies show that students may learn from their peers if they choose to cooperate or engage in joint production (Kimbrough et al. 2022). When the number of interactions among peers is sufficiently large, students move away from isolation and choose either joint production or mutual insurance as a mode of social interaction (De Giorgi & Pellizzari 2014). Third, changes in the network of friendships can also affect students' behavior. The change in the pool of potential friends can affect students' disruptive behaviors as well as nonacademic interests or risky behaviors, which have an indirect effect on individual learning Wu et al. (forthcoming), Lavy et al. (2011), Carrell et al. (2018).

Teachers can also respond to the changes introduced in the composition of their classes. Teachers may adjust the target student during lecture time and practice activities. For instance, more homogeneous groups can help teachers better target a larger share of students. In turn, heterogeneous groups may pose a challenge as targeting a specific performance level will require the teacher to dedicate additional time to cater to the needs of students too far away from the target. In addition, teachers can adjust their teaching practices and class-management strategies. The extent of these changes will of course depend on teachers' incentives to exert effort both in the classroom and outside of it. Even though teachers in our setting were neither informed about the treatment assignment nor aided with the provision of tools or pointers, their daily interactions with the group can trigger teachers' responses to the specific classroom composition and to the endogenous changes in student behavior.

For instance, a large share of high-achieving students may induce teachers to foster team work, allowing lower-achieving students to practice new material with their better-prepared peers. Alternatively, a large share of low-achieving students may induce teachers to put in more effort or seek complementary inputs to help weaker students. Teachers can also react to changes in the level of disruptive behavior either by getting discouraged and giving up or by reallocating time and resources to manage the classroom.

Repeated interactions between students and teachers will generate feedback effects among peers and continuous adaptation by teachers. It is therefore not possible to isolate the role of each "channel" on individual performance. This limitation is due to the nature of peer effects and is not unique to our research design. Similar to previous studies (Booij et al. 2017, Carrell et al. 2013, Duflo et al. 2011), we estimate the reduced-form causal impact of tracking and bimodal student allocations on performance and behavioral outcomes. These treatment effects are policy relevant for studying what the impact of these interventions would be at scale.

Tracking students by initial performance minimizes achievement differences of classmates in the same group. This can, in principle, trigger two possible changes. First, if there are positive peer effects that are linear in means, low-achieving students would no longer benefit from sharing the classroom with high-achieving students, but the high-achieving students would benefit even more than previously due to a higher concentration of similar high-achieving peers in their group. Absent other behavioral responses, tracking would lead to performance losses for low-achieving students and to performance gains for high-achieving students. Second, more homogeneous classes would allow teachers to better meet the needs of each classroom, thus allowing students in the extremes of the performance distribution to benefit from the tracking allocation. Thus, while high-achieving students are likely to benefit from tracking, the net effect of tracking on low-achieving students will depend on the relative magnitude of the positive effects of achievement-specific instruction and the negative peer effects.

Bimodal classrooms are advocated on the basis that they foster cooperative learning between students. It proposes classroom configurations that are likely to generate interactions between students with high and low initial performance levels. If students learn as a team, the best students in the class can have a guiding role, promoting effective learning. This model allows the teacher to generate different instructional scenarios that seek to adapt the curriculum to the needs of the group (Johnson et al. 1999). However, teachers also face the challenges of a diverse classroom, which may be more prone to disruptions and more difficult to manage. Moreover, the bimodal configuration exposes students to a larger-than-normal share of high-achieving peers, but it proportionally increases the share of low-achieving peers, which generates tension in terms of the externalities of diverse peers. Finally, students may decide to segregate based on performance within group, limiting the gains from cooperation with diverse peers. In the end, the direction and magnitude of the impact of bimodal classes on individual performance hinges on the ability of teachers to adapt to the classroom composition, the net effect of increasing exposure to both high and low achievers, and potential in-class segregation patterns.

The rest of this section provides empirical evidence on the behavioral responses of teachers and students under both classroom ability-groupings.

4.2 Changes in Teachers' Behavior

Despite the absence of pointers or guidance to adapt their behavior to the classroom they faced, teachers reacted in ways that aligned to the comparative advantage of each students allocation model. Panel A in Table 2 focuses on teachers' time management with respect to three main activities: dealing with disruptions in class, working on practice and feedback, and determining the extent of time devoted to lectures. Teachers working in tracking classrooms seized the homogeneity of the classes by allocating relatively more time to lectures, while reducing the time spent in practice and feedback activities among their students (see column [1]). By contrast, teachers facing bimodal classes tended to limit their role in the learning process by decreasing lecture time and prioritizing interactions among students by devoting more time to practice and feedback during class (see column [2]). This strategy also seems to have worked well to reduce disruptions in the class, which is remarkable given the higher level of heterogeneity in bimodal classes relative to the control group.

Columns [4]-[6] show the change in teachers' time management depending on the initial performance level of each tracked classroom. In medium- and high-performing groups, there was a statistically significant increase in the share of time allocated to lectures. However, low-performing tracking groups seemed to have proven more difficult to manage; the only significant change identified in these groups was an increase in the time teachers spent dealing with disruptions.

Despite the teachers' adaptation to the new classroom environment, neither those teaching tracking classes nor those teaching in bimodal classes changed their effort levels (see panel B in Table 2). On average, and irrespective of the treatment arm, there was no change in the time teachers spent helping students outside the classroom or on their self-reported level of teaching efficacy (their self-perceived ability to help their students learn). Teachers do report being relatively more effective in medium- and high-performing tracked classes. This could be related to teachers' satisfaction with the time-management changes they introduced when dealing with higher-performing tracking groups or, as we show next, to their ability to progress more smoothly through the class material when facing more homogeneous groups. Because more disruptions occurred in low-performing tracking groups, teachers reported that they were less effective at improving their students' academic achievement, although this effect is not statistically significant.

Table 2: Average Treatment Effects on Teacher Behavior

				Tracking by classroom type			
Baseline variable	Tracking	Bimodal	Obs.	Low	Medium	High	Obs.
	[1]	[2]	[3]	[4]	[5]	[6]	[7]
Panel A: Time management (% class)							
Disruptions	0.314	-1.867	514	1.619	-1.407	-0.509	347
	[0.633]	[0.667]***		[0.965]*	[0.939]	[1.172]	
Practice and feedback	-2.786	3.732	514	-0.666	-3.561	-3.121	347
	[1.352]**	[1.349]***		[2.086]	[2.212]	[2.356]	
Lecture	2.472	-1.864	514	-0.953	4.967	3.630	347
	[1.224]**	[1.134]		[1.851]	[2.134]**	[2.157]*	
Panel B: Effort							
Log(hours) prep. class	-0.080	-0.048	514	-0.043	-0.113	-0.052	347
	[0.075]	[0.082]		[0.096]	[0.103]	[0.109]	
Teaching efficacy (std.)	0.153	0.157	502	-0.062	0.346	0.287	340
	[0.097]	[0.104]		[0.139]	[0.155]**	[0.162]*	
Panel C: Teaching Practices							
Individual targeting (std.)	0.054	0.035	507	0.096	0.127	-0.095	341
	[0.100]	[0.109]		[0.131]	[0.159]	[0.187]	
Induced competition (std.)	-0.002	-0.196	505	-0.017	-0.105	0.005	339
	[0.104]	[0.119]*		[0.142]	[0.145]	[0.178]	
% of topics already covered	10.588	6.239	514	9.067	11.464	10.983	347
	[3.912]***	[4.294]		[4.101]**	[4.264]***	[4.804]**	

Note: Columns [1] and [2] report estimates of tracking and bimodal treatment effects using equation (1) on each outcome shown in each row. Columns [4]-[6] report estimates on tracking treatment effects by tracked classroom type. These estimates rely on tracking and control classrooms only: we estimated a model similar to equation (1) where in interact the tracking treatment-indicator variable (D_{sq}^{T}) with indicators for low-, medium-, and high-achievement tracked classes. Panel A focuses on variables related to the teachers' time management: percent of the class spent dealing with *disruptions*, doing *practice and giving feedback* or giving a *lecture*. Panel B focuses on the teachers' effort variables: *log(hours)* spent outside the class dedicated to preparing classes or grading and a measure of *teaching efficacy* is a standardized scale factor which captures the subjective perception of teachers to affect students' learning. Panel C focus on variables related to teaching practices: *individual targeting* is a standardized scale index variable which captures whether teachers pursue strategies to teach at the right level. The variable *induced competition* is a standardized scale index whether teachers follow pedagogical strategies that lead to more competition among students. *% topics covered* captures what percentage of the annual curriculum the teacher was able to cover by the end of the year. Section 3 of Appendix A (Data) provides more details regarding the construction of these variables. Estimators include strata fixed effects. Statistically significant at * 10%, ** 5% or *** 1% level. Standard errors clustered at the school level.

The last set of outcomes in Panel C refers to teaching practices. Both tracking and bimodal groups were able to cover a larger share of the curriculum. On average, teachers in tracking schools covered 10.6 percentage points more of the curriculum relative to control schools (see column [1]). This effect is quite homogeneous across tracking classes (see columns [4]-[6]), and suggests that there are similar gains from targeted instruction, irrespective of the average performance of students in a given classroom. Teachers in bimodal groups fostered a more cooperative environment by inducing less competition in the classroom. In more diverse classrooms, competition can activate rank concerns (Tincani 2017) and limit aspirations (Genicot & Ray 2020), particularly among bottom students who are likely to give up when half of the classroom is high performing. Teachers in bimodal classes seemed to have recognized this and were less likely to foster a competitive environment relative to control classes, perhaps in an attempt to keep low-performing students motivated.

All in all, teachers' responses matched quite well the pedagogical changes required to foster learning under each group formation strategy. Interestingly, while teachers' experience did not matter in tracking classes, teachers' ability to react to bimodal classes seemed to be related to how much experience they had. In fact, we find that only more seasoned teachers were able to extract significant average performance gains in bimodal classes.[20]

4.3 Changes in Students' Behavior

Table 3 shows intention-to-treat effects on absenteeism, academic effort, and risky behaviors. Panel A shows that absenteeism was on average unaffected in tracking groups (Column [1]). Column [2] reveals that students in bimodal schools, on the other hand, reduced their truancy by 1.6 percentage points and the number of absences (measured using administrative records) by 14 percent. Low-achieving students in tracking and bimodal classes exhibited similarly large and significant reductions in their number of absences per quarter (see columns [4]-[5]). These effects are comparable to those exhibited by high-achieving students in bimodal schools, presented in column [8]. This suggests that heterogeneous classes motivated low-ability students to come to classes relatively more.

Panel B focuses on effort measures. Column [2] shows that bimodal groups recorded a slight decrease in the number of hours students dedicated to study and work on their homework outside the classroom, which may correspond to a partial substitution effect of their greater effort levels while in the classroom. Focusing on the impacts on the effort index, we identify differential patterns by students' initial performance in tracking and bimodal schools. On one hand, students in high-performing tracking classrooms seem to have increased their effort in class, while students in low- and middle-performing classrooms do not seem to have reacted. In turn, the treatment effects on effort in bimodal groups are particularly large and strong among the high-achieving students. In fact, the change in their effort levels is 2.6 times higher relative to their counterparts in tracking classes.

[20] Table 1 in the Appendix explores the heterogeneity of treatment effects by teachers' experience and other dimensions such as school performance and students' characteristics.

Table 3: Average Treatment Effects on Student Behavior

				By Initial Achievement				
				Low		Medium	High	
Baseline variable	Tracking	Bimodal	Obs.	Tracking	Bimodal	Tracking	Tracking	Bimodal
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
Panel A: Absenteeism								
Truant	–0.002	–0.016	17,746	0.019	–0.027	–0.019	–0.010	–0.007
	[0.008]	[0.009]*		[0.012]	[0.013]**	[0.011]*	[0.010]	[0.010]
Avg. absences per quarter (log)	–0.086	–0.145	20,175	–0.166	–0.150	–0.082	–0.017	–0.149
	[0.087]	[0.074]*		[0.102]	[0.084]*	[0.100]	[0.086]	[0.069]**
Panel B: Effort								
Weekly hours of study (log)	0.014	–0.045	18,490	0.004	–0.027	0.015	0.032	–0.037
	[0.020]	[0.025]*		[0.037]	[0.038]	[0.026]	[0.027]	[0.032]
Effort index	0.059	0.135	13,917	0.109	0.061	–0.019	0.106	0.265
	[0.044]	[0.051]***		[0.074]	[0.082]	[0.064]	[0.059]*	[0.057]***
Panel C: Risky Behavior								
Bad behavior	0.005	–0.019	17,305	0.017	–0.014	–0.002	0.009	0.005
	[0.010]	[0.012]		[0.015]	[0.016]	[0.015]	[0.014]	[0.014]
Smoked or consumed alcohol	–0.008	–0.005	17,531	–0.009	0.001	–0.005	–0.010	–0.007
	[0.012]	[0.013]		[0.016]	[0.016]	[0.015]	[0.017]	[0.016]

Note: Columns [1] and [2] report estimates of tracking and bimodal treatment effects using equation (1) on each row outcome. Columns [4]-[8] report estimates on tracking and bimodal treatment effects by splitting the sample according to the student's initial achievement (low/medium/high refers to tercile 1/2/3 of the pre-treatment test score school distribution). We did not collect information for students in the medium tercile in bimodal schools. Panel A focus on student's absenteeism. Panel B focus on student's effort. Panel C focus on student's risky behavior. Estimators include strata fixed effects. All estimates correspond to equations that control for initial test score, an indicator if the exam score is equal to zero, having the initial score imputed, age, gender, an indicator of whether the student has pre-registration, administrative unit, number of classrooms per school and number of students by group. *Truant* is an indicator variable equal to one if the student reports skipping a class in the last two weeks without parents' permission. *Avg. absences per quarter* records the log of the absences recorded in administrative records. The variable *effort index* is an index comprised of three standardized scales that displays the student's classroom effort, classroom competition and disruptive behavior. *Bad behavior* is an indicator variable if the student reports having a fight or being suspended in the school year. *Smoked or consumed alcohol* is an indicator variable equal to one if the student reports ever having smoked cigarette or consumed alcohol. See Section 3 of Appendix A (Data) for details on the construction of these variables. Statistically significant at * 10%, ** 5% or *** 1% level. Standard errors are clustered at the school level.

In sum, we find that bimodal classes foster greater levels of effort and reduce the number of absences. This suggests that more heterogeneous groups can foster greater class-engagement levels among students. This effect is particularly important for high achievers —an insight that is aligned with their performance improvements and a higher probability of graduating on time. Despite increased effort levels among high-achieving peers and the changes in teachers' time management and strategies to adapt to the bimodal classroom environment, low-performing students in heterogeneous groups did not learn more than their peers in the control group. The performance gains from more diverse classes were limited even when the right endogenous responses from teachers and students were observed. This suggests the need to complement these group-allocation models with complementary interventions that better support teachers when dealing with these classes.

Notice that tracking fostered similar changes in effort at both extremes of the initial achievement distribution (0.11 of a standard deviation)– even if the effect among the lowest achievers was not statistically significant. Because teachers had a hard time adapting to low-performing classes, with more time allocated to deal with disruptions and a decrease in teaching efficacy, the higher effort levels exerted by low-performing tracking students were not enough to yield learning gains relative to the control.

Panel C of Table 3 presents the results of both group-allocation strategies on self-reported risky behaviors such as violence or extreme disruptive behavior and substance abuse (i.e., smoking or drinking alcohol). During adolescence, youth start developing their own identities. They become more internally directed, and the influence of parents and teachers is overshadowed by the role of peers. Adolescence is also characterized by a high amount of experimentation and search for new experiences, which increases exposure to risky behaviors. Engaging in risky behavior has potential effects on immediate school performance as well as on future educational and labor-market trajectories. Indeed, Carrell et al. (2018) show that disruptive peers (i.e., someone exposed to domestic violence) have long-term consequences on labor-market outcomes: exposure to a disruptive peer during elementary school reduces earnings by age 24 to 28 by 3 percent. Other studies focus on the role of peers in adolescent decision-making and non-cognitive outcomes. For instance, Card & Giuliano (2013) identify significant peer effects in sexual initiation, smoking, marijuana use, and truancy. Lavy & Sand (2012) show that the impact of the number of friends on students' educational outcomes is partly mediated by noncognitive traits such as cooperative behavior, violent behavior and social satisfaction. Wu et al. (forthcoming) find that having a deskmate with high levels of extraversion and agreeableness fosters extraversion and agreeableness.

Our results contrast with those found in previous studies. We focus on observed behaviors —which may reflect intermediate noncognitive outcomes. We fail to identify any significant impact of either group-allocation model on risky behaviors, irrespective of the initial performance level of the student. We do not find evidence to suggest that the absence of performance gains among low achievers in the tracking and bimodal models was due to increased risky behaviors.

4.4 Student Sorting and Changes to Friendship Network

Students in bimodal classroom could have made friendships with students of similar pre-treatment academic performance, thus limiting the extent of cooperation with diverse peers. Fortunately, our survey at the end of seventh grade collected data on students' friendship networks that allows us to test for homophily. Specifically, we asked each student to report her top three friends as well as her best friend in the classroom.

Table 4 reports the treatment effects on endogenous sorting into friendships among low-performing students in bimodal schools. For each low-achieving student we estimate the fraction of friends that belong to each academic achieving tercile. The first column reports the difference in the percentage of friends from a given tercile relative to the control group. In general, low achieving students were more likely to befriend low achievers and high achievers than they would have otherwise befriended had they been allocated to a control group classroom. Row 1 in column [1] indicates that low performing students in bimodal classes are 17 percentage points more likely than their counterparts in the control group to have friends from their own performance tercile. The second row under column [1] reports that a similar treatment effect is identified for friendships with high performing peers: low-achieving students from bimodal classes are 18 percentage points more likely than control group students to befriend students from the highest tercile. Rows 3 and 4 in column [1] show that very similar effects are identified if we only focus on the best friend reported by each student. At first glance, the degree of segregation that we identify in our setting is much smaller compared to that quantified in Carrell et al. (2013).[21]

Table 4: Treatment Effects on Friend Choices Among Low-Performing Students

Treatment effect on:	Actual peers (se) [1]	Random peers (sd) [2]	Actual minus random $P(A < R)$ [3]
% of low-achieving friends	0.170	0.148	0.022
	[0.014]***	[0.012]***	0.030
% of high-achieving friends	0.182	0.186	-0.004
	[0.016]***	[0.012]***	0.619
Best friend: low achiever	0.188	0.148	0.039
	[0.016]***	[0.019]***	0.023*
Best friend: high achiever	0.179	0.185	-0.006
	[0.020]***	[0.020]***	0.607

Note: Columns [1] reports estimates and standard errors of actual bimodal treatment effects among low-performing students on each row outcome. Column [2] reports the average and standard deviation of the same estimated coefficient using 3,000 iterations of resampled friend random assignments within each classroom. Column [3] reports the difference between coefficients in columns [1] and [2] and the p-values of the null hypothesis that the coefficients of random friendships are equal to those from the actual friendships. Estimators include strata fixed effects. Sample restricted to bimodal and control schools. Statistically significant at * 10%, ** 5% or *** 1% level. Standard errors are clustered at the school level.

These changes in individual networks relative to control classrooms are, to some extent, expected. They are partly explained by the more diverse pool of peers that bimodal classes offer when compared to control groups. Low-performing students in bimodal classes can only befriend low- and high- achieving students because middle-achieving students were tracked out. It is only natural to ask how much of this effect is explained by classroom composition vis-à-vis preferences when choosing friends. To isolate the effect of classroom composition,

[21] See Table VIII in Carrell et al. (2013).

we randomly allocated students to peers in their classroom and labeled them as friends. We choose as many friends as each student reported in the survey with replacement, thus allowing for the existence of complex networks with unidirectional links. We repeat these random allocations 3,000 times in each classroom and report the average treatment impacts across all iterations in column [2].

Column [3] presents the difference between the observed proportion of friends and a random friendship network, capturing endogenous biases when making friends. The results in rows 2 and 4 show that virtually all of the increase in links with high-achieving students is explained by changes in the pool of potential friends. Low-achieving students in bimodal classes befriended more high-performing students, but only to the extent that the latter were more abundant. In turn, some, but not all of the observed increase in low-achieving students' links with peers from similar achievement level is explained by the change in the pool of students. Low-achieving students are 2.2 percentage points more likely to befriend similar peers than what is predicted by the classroom composition. The results thus suggest that there is a slight increase in homophily among low achievers.

These results indicate that the degree of exposure to high-performing peers was limited and that low-performing students had a tendency to befriend similar peers at a higher rate. This effect is modest, but it could still limit positive spillovers from high-achieving peers.

5 Conclusions

The learning challenges faced during adolescence are quite different from those faced during childhood. At this older stage of development, youth become more internally directed, and the influence of parents is overshadowed by the role of peers. Middle school is often the last mandatory stage in the formal education system in many developing countries and, as such, it constitutes one of the last chances to systematically address adolescents' skills deficiencies. The extensive literature on peer effects in the educational setting confirms the large impact of classmates on both individual academic performance and on non-academic outcomes. Surprisingly, less work has been done to study the impact of controlled classroom-grouping strategies based on initial academic performance of middle-school students. These interventions do not require extra resources and may thus produce cost-effective gains in learning while being easily scalable interventions.

By conducting a large randomized controlled trial in public middle schools in Mexico City, we are able to analyze the learning effects of two different ability-grouping models: tracking and bimodal classrooms. We identify important and similar average gains in student performance in both types of allocations. At the end of seventh grade, after one year of exposure, both grouping strategies yield similar average performance gains of about 0.08 of a standard deviation, with greater and more persistent effects among students who initially showed high performance levels. Though low-performing students did not benefit, there were no performance losses among them.

Even though both grouping models resulted in similar average performance gains, the magnitude of the effects diverges across models once we look at the differential impacts across students of different initial academic performance. The largest learning gains were accrued by students in the tracking classes for the top achievers, followed by the top students in bimodal classroom and medium-performing students in tracking classrooms. This ranking of effect sizes is reasonable. The top tracking class had everything going for it: a concentration of high-performing peers and a very homogeneous classroom that facilitated the work of the teacher. The relatively smaller gains obtained by high-performing peers in bimodal classes suggest that a more cooperative interaction model was not enough to compensate for reduced levels of exposure to high-performing peers (relative to the exposure experienced by top tracking classes) or for the greater challenges faced by teachers when tailoring instruction to a more diverse group. Finally, the modest learning gains obtained by medium performers under tracking confirm that exposure to high-performing peers is important, and that classroom homogeneity is only partially effective to foster learning.

Low-achieving students did not experience performance gains under either grouping strategy. In the case of tracking, this suggests that teachers were able to neutralize the potential negative influence of a higher concentration of lower-performing peers. In the case of bimodal schools, this result can be driven by multiple factors. First, even if bimodal classrooms offered low-performing students the chance to interact with a higher share of high-performing peers relative to the control, they also exposed them to more students of their same low level of initial performance. Second, the extent of spillovers from high-achieving students was limited by an increased propensity to befriend other low-achieving students.

Our estimated treatment effects capture the impact of tracking and bimodal classrooms in a situation that did not incorporate any additional, complementary investments. In other words, our causal impacts isolate the effect of changing the classroom composition, allowing for endogenous responses of both students and teachers. Our results highlight a greater potential of tracking to foster learning among top achievers in middle school, without hurting lower-performing students. While tracking alone might not be enough to improve learning along the complete distribution of students' skills, there is a potential set of complementary, low-cost investments that could foster additional performance improvements among low-achievers.

References

Alvarez-Marinelli, H., Berlinski, S. & Busso, M. (2021), 'Remedial education: Evidence from a sequence of experiments in colombia', *Journal of Human Resources* .

Balestra, S., Sallin, A. & Wolter, S. (2021), 'High-ability influencers? the heterogeneous effects of gifted classmates', *Journal of Human Resources* .

Betts, J. R. (2011), The economics of tracking in education, *in* 'Handbook of the Economics of Education', Vol. 3, Elsevier, pp. 341–381.

Betts, J. R. & Shkolnik, J. L. (2000), 'Key difficulties in identifying the effects of ability grouping on student achievement', *Economics of Education Review* **19**(1), 21–26.

Booij, A., Leuven, E. & Oosterbeek, H. (2017), 'Ability peer effects in university: Evidence from a randomized experiment', *Review of Economic Studies* **84**, 547–578.

Bramoullé, Y., Djebbari, H. & Fortin, B. (2009), 'Identification of peer effects through social networks', *Journal of Econometrics* **150**(1), 41–55.

Busso, M. & Frisancho, V. (2021), 'Good peers have asymmetric gendered effects on female educational outcomes: Experimental evidence from mexico', *Journal of Economic Behavior and Organization* **189**, 727–747.

Calsamiglia, C. & Güell, M. (2018), 'Priorities in school choice: The case of the boston mechanism in barcelona', *Journal of Public Economics* **163**, 20–36.

Card, D. & Giuliano, L. (2013), 'Peer Effects and Multiple Equilibria in the Risky Behavior of Friends', *The Review of Economics and Statistics* **95**(4), 1130–1149.

Carrell, S., Fullerton, R. & West, J. (2009), 'Does your cohort matter? measuring peer effects in college achievement', *Journal of Labor Economics* **27**(3), 439–464.

Carrell, S., Hoekstra, M. & Kuka, E. (2018), 'The long-run effects of disruptive peers', *American Economic Review* **108**(11), 3377–3415.

Carrell, S., Sacerdote, B. & West, J. (2013), 'From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation', *Econometrica* **81**(3), 855–882.

Cools, A., Fernández, R. & Patacchini, E. (2021), The asymmetric gender effects of high flyers, Technical report.

De Giorgi, G. & Pellizzari, M. (2014), 'Understanding social interactions: Evidence from the classroom', *Economic Journal* **124**(579), 917–953.

De Giorgi, G., Pellizzari, M. & Redaelli, S. (2010), 'Identification of social interactions through partially overlapping peer groups', *American Economic Journal: Applied Economics* (2).

Delaney, J. M. & Devereux, P. J. (2022), Rank effects in education: What do we know so far?, Technical report, IZA Discussion Papers, No. 15128.

Duflo, E., Dupas, P. & Kremer, M. (2011), 'Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in kenya', *American Economic Review* **101**, 1739–1774.

Evans, D. & Popova, A. (2015), What really works to improve learning in developing countries? an analysis of divergent findings in systematic reviews, Technical report, World Bank Policy Research Working Paper No. 7203.

Genicot, G. & Ray, D. (2020), 'Aspirations and economic behavior', *Annual Review of Economics* **12**(1), 715–746.

Hart, S., Stewart, K. & Jimerson, S. R. (2011), 'The student engagement in schools questionnaire (sesq) and the teacher engagement report form-new (terf-n): Examining the preliminary evidence', *Contemporary School Psychology* **15**(1), 67–79.

Imberman, S., Kugler, A. & Sacerdote, B. (2012), 'Katrina's children: Evidence on the structure of peer effects from hurricane evacuees', **102**.

Johnson, D., Johnson, R. & Anderson, D. (1983), 'Social interdependence and classroom climate', *The Journal of Psychology* **114**, 135–142.

Johnson, D., Johnson, R. & Holubec, E. J. (1999), *El aprendizaje cooperativo en el aula*, Paidós.

Kimbrough, E. O., McGee, A. D. & Shigeoka, H. (2022), 'How do peers impact learning? an experimental investigation of peer-to-peer teaching and ability tracking', *Journal of Human Resources* **57**(1), 304–339.

Lavy, V., Paserman, D. & Schlosser, A. (2011), 'Inside the black box of ability peer effects: Evidence from variation in the proportion of low achievers in the classroom', *The Economic Journal* **122**, 208–237.

Lavy, V. & Sand, E. (2012), The friends factor: How students' social networks affect their academic achievement and well-being?, Nber working papers, National Bureau of Economic Research, Inc.

Lavy, V. & Schlosser, A. (2011), 'Mechanisms and impacts of gender peer effects at school', *American Economic Journal: Applied Economics* **3**(2), 1–33.

Manski, C. F. (1993), 'Identification of endogenous social effects: The reflection problem', *The Review of Economic Studies* **60**(3), 531–542.

Martin, W. H. (1927), 'The results of homogeneous grouping in the junior high school', *Unpublished doctoral dissertation, Yale University* .

Midgley, C., Maehr, M., Hruda, L., Anderman, E., Anderman, L., Freeman, K., Gheen, M., Kaplan, A., Kumar, R., Middleton, M., Nelson, J., Roeser, R. & Urdan, T. (2000), Manual for the patterns of adaptive learning scales, Technical report, University of Michigan.

Muralidharan, K., Singh, A. & Ganimian, A. J. (2019), 'Disrupting education? experimental evidence on technology-aided instruction in india', *American Economic Review* **109**(4), 1426–60.

Murphy, R. & Weinhardt, F. (2020), 'Top of the Class: The Importance of Ordinal Rank', *The Review of Economic Studies* **87**(6), 2777–2826.

Slavin, R. E. (1987), 'Ability grouping and student achievement in elementary schools: A best-evidence synthesis', *Review of educational research* **57**(3), 293–336.

Slavin, R. E. (1990), 'Achievement effects of ability grouping in secondary schools: A best-evidence synthesis', *Review of educational research* **60**(3), 471–499.

Tincani, M. (2017), 'Heterogeneous peer effects and rank concerns: Theory and evidence'.

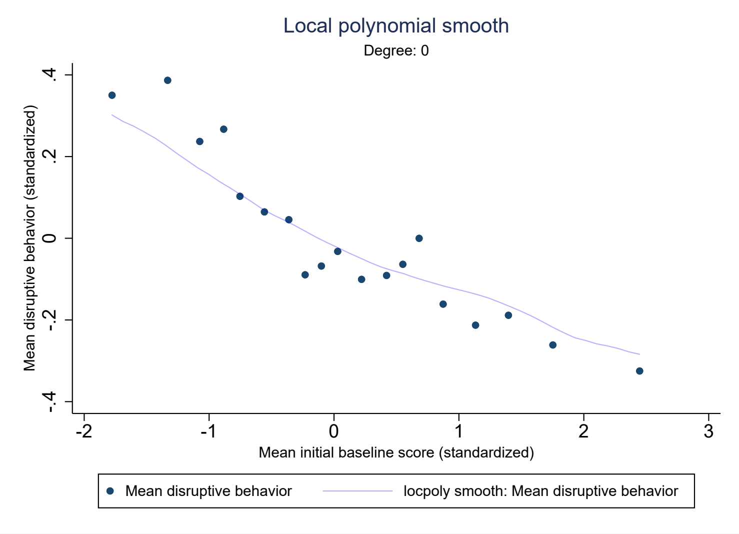
Tincani, M. (2018), Heterogeneous peer effects in the classroom, Technical report.

Turney, A. H. (1931), 'The status of ability grouping', *Educational Administration and Supervision* **17**(2), 110–27.

Wu, J., Zhang, J. & Wang, C. (forthcoming), 'Student performance, peer effects, and friend networks: Evidence from a randomized peer intervention', *American Economic Journal: Economic Policy* .

Appendix: Additional Figures and Tables

Appendix Figure 1: Mean disruption and mean baseline test score (control group only)



Appendix Table 1: Treatment Effect: Heterogeneity

Sample	Treatment Effect by:		Tracking [1]	Bimodal [2]	Observations [3]
Panel A:					
Teacher	Experience	High	0.085 [0.032]***	0.123 [0.033]***	12849
		Low	0.077 [0.035]**	0.038 [0.031]	14438
		p-value of differences between groups	0.846	0.026	
Panel B:					
Student	Sex	Boys	0.080 [0.031]***	0.081 [0.038]**	9727
		Girls	0.067 [0.032]**	0.078 [0.036]**	9399
		p-value of differences between groups	0.595	0.937	
	Socio-economic Status	Low	0.062 [0.027]**	0.090 [0.032]***	8880
		High	0.090 [0.034]***	0.071 [0.040]*	10246
		p-value of differences between groups	0.220	0.492	

Note: Columns [1] and [2] report estimates of tracking and bimodal treatment effects using equation (1) on aggregate standardized test score by each row heterogeneity. Sample is splitted for each subgroup. Panel A analyzes the differential treatment effect by teacher's experience. Panel B analyzes the treatment effect by student's characteristics. We did not collect information for students in the medium tercile in bimodal schools.. Estimators include strata fixed effects, where strata are sets of 3 schools each sharing common characteristics. All estimates correspond to equations that control for initial test score, an indicator if the exam score is equal to zero, having the initial score imputed, age, gender, an indicator of whether the student has pre-registration, administrative unit, number of classrooms per school and number of students by group. Student's socioeconomic status is measured as high if both the students' parents have secondary education or higher, and low otherwise.* statistically significant at 10%; ** at 5%; *** at 1% level. Standard errors clustered at the school level.

A Data Appendix

1 Data Sources

Student administrative data

Student administrative data come from the Public Education Secretary of Mexico (SPE for its acronym in Spanish), specifically from two sources: the Enrollment and Student Allocation System (Sistema Anticipado de Inscripción y Distribución, SAID) and the Data System Arturo Rosenblueth (Centro de Desarrollo Informatico Arturo Rosenblueth, CDIAR).

SAID data contain information from student registration forms which includes: student's identification variables; the school assigned before the beginning of the school year; baseline IDANIS test scores (more information on this in the next section); school preferences (three preferred schools for starting middle school); the school attended for primary school; whether the student attended kindergarten; whether they have special needs; whether the mom/father completed secondary school or more; student's sex; date of birth; weight; height; and household size.

CDIAR's end-of-the-year database has information on the students' GPA (total and/or per subject) and absenteeism. Bimonthly data provides the students' school and group location throughout the entire school year (one observation per bimester).

General administrative data

CDIAR school-level data has information on number of classrooms per school, the capacity of each classroom in 2015, and previous levels of enrollment. We also use the school average ENLACE score –a test that aims to assess the quality of the education given in Math and Language– as a school quality proxy. CDIAR also provided information on teacher variables such as age, sex, and academic diplomas.

Survey data

At the end of the first academic year, after the intervention took place, we administered student and teacher surveys, as well as the endline exam. Questionaries for students and teachers were specifically made to assess the impact of the program in different outcomes such as teacher and student behavior, as well as teacher and student effort. In tracking schools, all students answer the exam; in bimodal and control schools, three classrooms per school were randomly selected and all students in those classrooms took the test. The endline survey collected information on the students' friends (peers), students' and peers' behaviors, teachers' behavior and family daily routine are gathered. We construct with these variables several scales (this is further explained below).

2 IRT calculations for endline exams

Tables A.1 and A.2 show the IRT results using a two-parameter model and dividing samples between treatment and control. As can be seen, difficulty and discriminatory parameters between both samples are similar in both direction and significance. We take this as evidence that the test was similar for students in different treatment arms.

Appendix Table A.1: Grade 7 IRT

Subscore	Item	Difficulty Parameter - Control (1)	Discriminatory Parameter - Control (2)	Difficulty Parameter - Treatment (3)	Discriminatory Parameter - Treatment (4)
Literacy	1	0.886***	0.638***	0.755***	0.666***
	2	0.419***	0.846***	0.440***	0.828***
	3	-0.564***	0.762***	-0.654***	0.820***
	4	-1.210***	0.834***	-1.254***	0.808***
	5	-1.547***	1.802***	-1.598***	1.892***
	6	2.494***	0.317***	2.480***	0.332***
	7	1.022***	0.442***	0.711***	0.483***
	8	-1.564***	1.494***	-1.622***	1.486***
	9	-0.628***	0.563***	-0.659***	0.587***
	10	4.229***	0.157***	3.123***	0.201***
	11	-0.157***	0.920***	-0.290***	0.885***
	12	1.859***	0.501***	1.861***	0.494***
	13	-0.312***	1.491***	-0.353***	1.569***
	14	-1.070***	0.823***	-1.146***	0.771***
	15	-5.865***	-0.207***	-6.139***	-0.199***
	16	-1.242***	1.915***	-1.382***	1.805***
	17	1.918***	0.592***	1.586***	0.682***
	18	-1.251***	2.332***	-1.250***	2.341***
	19	-0.946***	1.908***	-0.976***	2.077***
	20	-1.075***	1.851***	-1.168***	1.710***
	21	-1.082***	2.575***	-1.123***	2.506***
	22	-0.953***	1.436***	-0.985***	1.503***
	23	2.808***	0.449***	2.219***	0.508***
	24	-1.042***	1.869***	-1.067***	1.997***
Math	25	4.843***	0.353***	6.788***	0.237***
	26	0.079	-1.001***	0.184***	-1.144***
	27	-6.572***	-0.142***	-3.693***	-0.240***
	28	-5.181	-0.079	-2.209***	-0.194***
	29	0.146***	-0.656***	0.096***	-0.706***
	30	-16.067	-0.080	-6.218***	-0.218***
	31	-1.208***	-0.935***	-1.036***	-1.018***
	32	-1.147***	-1.051***	-0.955***	-1.175***
	33	-0.026	-1.101***	0.026	-1.145***
	34	-1.430***	-0.804***	-1.222***	-0.864***
	35	-0.376***	-0.657***	-0.303***	-0.695***
	36	-0.731***	-0.520***	-0.568***	-0.555***
	37	-5.527***	-0.219***	-3.782***	-0.305***
	38	-3.899***	-0.212***	-2.422***	-0.335***
	39	-4.019***	-0.257***	-3.568***	-0.276***
	40	-3.131***	-0.476***	-2.208***	-0.624***
	41	-2.682***	-0.381***	-2.199***	-0.460***
	42	-0.298***	-0.966***	-0.255***	-1.028***
	43	0.065	-0.403***	0.144***	-0.490***
	44	5.333***	0.233***	11.574***	0.105***
	45	-2.561***	-0.782***	-2.073***	-0.920***
	46	-0.557***	-0.559***	-0.422***	-0.635***
	47	-2.348***	-0.386***	-1.925***	-0.480***
	48	-10.903***	-0.104***	-7.561***	-0.145***
Abstract	49	-0.984***	-0.778***	-0.802***	-0.862***
	50	-0.501***	-1.071***	-0.492***	-1.046***
	51	-0.456***	-0.793***	-0.386***	-0.919***
	52	-0.607***	-0.740***	-0.509***	-0.827***
	53	-0.479***	-0.772***	-0.415***	-0.833***
	54	-0.758***	-0.649***	-0.690***	-0.763***
	55	-1.077***	-0.720***	-0.997***	-0.755***
	56	-1.641***	-0.430***	-1.573***	-0.484***
	57	-0.634***	-0.769***	-0.474***	-0.861***
	58	-0.373***	-0.707***	-0.236***	-0.847***
	59	-0.966***	-0.504***	-0.791***	-0.594***

Note: IRT estimates using a two-parameter model. Significance levels * 10%, ** 5%, *** 1%.

Appendix Table A.2: Grade 8 exam IRT

Subscore	Item	Difficulty Parameter - Control	Discriminatory Parameter - Control	Difficulty Parameter - Treatment	Discriminatory Parameter - Treatment
		(1)	(2)	(3)	(4)
Literacy	1	-3.473***	0.505***	-3.576***	0.525***
	2	-2.633***	0.805***	-2.669***	0.825***
	3	-3.142***	1.005***	-3.245***	1.005***
	4	3.575***	0.559***	4.733***	0.429***
	5	-0.408***	0.663***	-0.382***	0.675***
	6	7.527***	0.307***	5.747***	0.403***
	7	-1.515***	1.380***	-1.616***	1.361***
	8	-0.907***	0.692***	-0.898***	0.742***
	9	-1.473***	1.359***	-1.562***	1.370***
	10	-0.689***	0.839***	-0.786***	0.791***
	11	-0.636***	0.783***	-0.602***	0.894***
	12	-0.708***	0.569***	-0.710***	0.650***
	13	-1.745***	0.518***	-1.899***	0.475***
	14	-0.865***	1.212***	-0.912***	1.310***
	15	1.555***	0.581***	1.600***	0.516***
	16	-0.958***	0.778***	-1.108***	0.738***
	17	0.095	0.595***	-0.144***	0.623***
	18	-0.794***	1.882***	-0.856***	1.710***
	19	1.180***	0.340***	0.912***	0.408***
	20	-0.918***	1.540***	-0.979***	1.463***
	21	-0.835***	0.408***	-0.763***	0.401***
	22	1.485***	0.551***	1.279***	0.572***
	23	0.536***	0.479***	0.678***	0.426***
	24	-0.837***	0.357***	-0.773***	0.410***
Math	25	-0.502***	0.922***	-0.572***	0.898***
	26	4.190***	0.164***	6.874***	0.102***
	27	-0.212***	0.514***	-0.223***	0.541***
	28	-0.240***	0.632***	-0.277***	0.644***
	29	0.239***	0.846***	0.243***	0.880***
	30	3.138***	0.411***	3.130***	0.404***
	31	105.604	0.020	28.302	0.075
	32	9.788***	0.113***	112.339	0.009
	33	16.095	0.079	9.419***	0.134***
	34	1.572***	0.815***	1.588***	0.756***
	35	2.191***	0.461***	1.950***	0.504***
	36	0.010	0.708***	0.004	0.706***
	37	0.978***	0.734***	0.854***	0.765***
	38	-0.565***	0.648***	-0.654***	0.575***
	39	0.657***	0.610***	0.654***	0.583***
	40	-0.612***	0.977***	-0.619***	1.033***
	41	-1.500***	0.350***	-1.514***	0.341***
	42	1.290***	0.309***	1.009***	0.345***
	43	0.332***	0.897***	0.314***	0.913***
	44	1.234***	0.620***	1.184***	0.622***
	45	2.241***	0.204***	1.991***	0.255***
	46	0.825***	0.394***	0.823***	0.364***
	47	-136.918	-0.008	11.190***	0.092***
	48	6.629***	0.178***	4.769***	0.232***
Abstract	49	-0.041	1.257***	-0.108***	1.337***
	50	0.206***	1.412***	0.118***	1.428***
	51	0.908***	0.751***	0.717***	0.790***
	52	-0.268***	1.947***	-0.272***	2.165***
	53	-2.705***	-0.616***	-2.615***	-0.641***
	54	0.091	1.083***	0.094***	1.181***
	55	0.338***	1.340***	0.289***	1.338***
	56	-0.059	1.721***	-0.069***	1.694***
	57	58.599	0.023	-55.765	-0.024
	58	0.939***	0.539***	0.816***	0.532***
	59	-3.025***	-0.566***	-2.940***	-0.602***

Note: IRT estimates using a two-parameter model. Significance levels * 10%, ** 5%, *** 1%.

Instrument psychometric properties

Endline test design was made following the same structure as the test carried out by the students before the beginning of middle school to define their middle school placements. Table A.3 shows the correlations between the baseline IDANIS test score and the main achievement outcomes.

Appendix Table A.3: Correlations several outcomes with baseline score

Ability measure	Correlation with Baseline IDANIS
IDANIS 7th Grade	0.611***
IDANIS 8th Grade	0.546***
GPA 9th Grade	0.323***
Math score - IDANIS 7th Grade	0.472***
Math score - IDANIS 8th Grade	0.436***
Math GPA 9th Grade	0.283***
Language score - IDANIS 7th Grade	0.519***
Language score - IDANIS 8th Grade	0.434***
Language GPA 9th Grade	0.259***
Abstract reasoning score - IDANIS 7th Grade	0.445***
Abstract reasoning score - IDANIS 8th Grade	0.426***

3 Variables and scales

Tables A.4, A.5, A.6 and A.7 describe the several variables used in the paper.

Appendix Table A.4: Variables and description - Balance variables

Variable	Description
	At the student level
Secondary or higher, father	Dummy variable that takes the value of one if the father finished secondary, at the least.
Male	Dummy variable that takes the value of one if the student is male.
Special needs	Dummy variable that takes the value of one if the student reports to have special needs.
Older than 12	Dummy variable that takes the value of one if the student is more than 12 years old at the beginning of the school year.
Initial test score	Initial IDANIS score of student. All students are divided in terms of their skills based on their performance in this exam.
Primary school GPA	GPA during primary; grades are between 6 and 10.
Secondary or higher, mother	Dummy variable that takes the value of one if the mother finished secondary, at the least.
Living with both parents	Dummy variable that takes the value of one if the student lives with both parents.
Proportion missing initial test score	Students who enrolled late in the education system do not have a grade in the Initial test score. This variable is at the student level and considers how many students are extemporaneous (enrolled late).
Accepted transfer	Authorized change of school (from that originally assigned).
	At the school level
Avg. test score (ENLACE) 2013	Average score in ENLACE exam in 2013 of the school.
SD initial test score	Standard deviation of the initial IDANIS score.
Change in # of students, 2012-2015	Change in number of students between 2012 and 2015
Number of classrooms	Average number of groups.
Class size	Average number of students per class.
	At the teacher level (sample does not comprise the complete teacher universe
Avg. age of teacher	Average age of teacher.
Gender	Proportion of female teachers.
Bachelor's degree	Proportion of teachers that own a bachelor's degree.
Years of experience in secondary level	Years of experience teaching in the secondary level.

Appendix Table A.5: Variables and description - Compliance and Student Churning variables

Variable	Description
Compliance and group stability variables	
Nominal compliance Baseline	Proportion of students in a classroom who are seated in the originally assigned classroom in the first bimester.
Group Stability 8th grade	Proportion of complier students in a classroom with respect to the original number of students seated in the classroom in bimester 1 of 7th grade.
Group Stability 9th grade	Proportion of complier students in a classroom with respect to the original number of students seated in the classroom in bimester 1 of 7th grade

Appendix Table A.6: Variables and description - Outcome variables

Variable	Description
Student outcomes	
Total score	Global exam score standardized relative to the control group. The total score is computed by adding up the three test sub-scores, say, math, literacy and abstract reasoning.
Math	Math exam score standardized to control group.
Literacy	Literacy exam score standardized to control group.
Abstract reasoning	Abstract reasoning exam score standardized to control group.
Total score (8th grade)	Global exam score standardized relative to the control group. The total score is computed by adding up the three test sub-scores, say, math, literacy and abstract reasoning. This exam was taken 2 years after beginning of intervention (2017).
Math (8th grade)	Math exam score standardized to control group. This exam was taken 2 years after beginning of intervention (2017).
Literacy (8th grade)	Literacy exam score standardized to control group. This exam was taken 2 years after beginning of intervention (2017).
Abstract reasoning (8th grade)	Abstract reasoning exam score standardized to control group. This exam was taken 2 years after beginning of intervention (2017).
GPA (9th grade)	Overall GPA at the end of the 2017-2018 academic year.
Graduation (9th grade)	Dichotomous variable equal to 1 if 3rd grade GPA is above or equal to 6 or 0 otherwise; if GPA is above or equal to 6, the student will graduate middle school.
Language GPA (9th grade)	Language GPA at the end of the 2017-2018 academic year.
Math GPA (9th grade)	Math GPA at the end of the 2017-2018 academic year.
Effort index	An index comprised of the standardized scales of Classroom effort, Classroom competition and Disruptive behavior. Specifically, it is done adding the scores in Classroom effort and Classroom competition and subtracting Disruptive behavior.
Classroom effort	An index done through factor analysis and standardized to control. It captures dimensions of the students' effort within the classroom, such as working hard, paying attention in classes, understanding difficult problems, time spent doing homeworks and participating in class activities.
Classroom competition	Depicts the students' goal of establishing their competence level. It is done through factor analysis and standardized to control. It considers components such as showing other students how good you are, demonstrate to others that class work is easy and that it is important for you to look smart in front of your peers.
Disruptive Behavior	Depicts the students' propensity to cause disruptions at school. It is done through factor analysis and standardized to control. It considers components such as following instructions, mocking teachers during class, having trouble with teachers during classes, disturbing the class, behaving in a way that annoys teachers.
Weekly hours of study	Log of the number of weekly hours students took to study or complete homework on mathematics and literacy outside the school.
Average absences per quater (log)	Logarithm of the average absences reported for the student per quarter. In case only one quarter was reported, this was the number used.
15 easiest items	Standardized score of the exam comprising the students' results of the 15 easiest items in the exam (according to the IRT difficulty parameter model). Standardized with respect to the whole population.
15 most difficult items	Standardized score of the exam comprising the students' results of the 15 most difficult items in the exam (according to the IRT difficulty parameter model). Standardized with respect to the whole population.
Teacher outcomes	
Time outside class	Log of the number of hours a week the teachers dedicate to prepare classes, grade exams or homeworks, complete administrative tasks and design activities focused on students with learning barriers.
Teaching efficacy	Measures the teachers' perspective that their effort and work can influence the students' performance and make a difference in their lives. It is constructed with factor analysis and standardized to control.
Individual targeting	Captures the strategies followed by the teachers to develop the idea of competence among students, for example, giving them the option to choose among several different activities, offer them a wide range of exercises according to their needs and abilities and acknowledge their improvements. The scale is constructed with factor analysis and standardized to control.
Induced competition	Captures the strategies followed by the teachers to demonstrate competence among students, for instance, highlighting the best students as a model for the rest, explaining how the performance of each student compares with one another and giving privileges to the students who make their work better. The scale is constructed with factor analysis and standardized to control.
Disruptions	Time spent dealing with disruptions to the class.
Practice and feedback	Time allocated to giving feedback to students.
Lecture	Time spent giving the lecture.
% of topics covered	Average percentage of covered topics from the mathematics and language curricula.

Appendix Table A.7: Variables and description - Heterogeneity exercises

Variable	Description
School variables	
Initial test score average	Divides the students in groups depending on whether the school they attend has an initial test score above or below school score average.
Class size average	Divides the students in groups depending on whether the school they attend has a class size that is above or below school class size average.
Teacher variables	
Experience	Divides groups into two, in the High experience case, teachers have above mean experience, Low stands for the opposite.
Student variables	
Gender	Divides students between males and females.
Socio-economic Status	Low is defined as both parents not having finished secondary, or either one not having finished it. High is defined as both parents having finished secondary or a higher level of education.

Scales

The following two tables describe the factor models used to compute each scale index measure. For each index we list: the highest eigenvalue, the Cronbach's Alpha coefficient, the items that compose the index, and the individual loading factors of each item.

Appendix Table A.8: Student scales - Items and Loadings

Scale	Items	Loadings
Classroom Effort	When I am given a difficult task, I make an effort to solve it	0.647
Eigenvalue: 2.421	I work as hard as I can in class	0.741
Cronbach's Alpha: 0.712	I participate in class activities	0.485
	I pay attention in class	0.600
	When I'm in class, I pretend to work	0.290
	I get distracted in classes	0.277
	At school, I only try the bare minimum	-0.200
	If I don't understand a topic, I work on it until I understand it	0.530
	I strive to do well in school	0.614
Classroom Competition	One of my goals is to show others that I am good	0.615
Eigenvalue: 1.764	For me it is important that others believe that I am good	0.768
Cronbach's Alpha: 0.753	Looking smarter than others	0.664
	Show that school work is easy for me	0.596
Disruptive Behavior	I follow the instructions of my teachers during class	-0.245
Eigenvalue: 2.250	I annoy my teachers during class	0.732
Cronbach's Alpha: 0.774	I have problems with my teachers during class	0.747
	I mess up during class	0.710
	I behave in a way that annoys my teachers in class	0.769

Appendix Table A.9: Teacher scales - Items and Loadings

Scale	Items	Loadings
Individual targeting	I offer students to choose between several activities	0.069
Eigenvalue: 1.290	I offer students a wide range of tasks	0.299
Cronbach's Alpha: 0.438	I make an effort to recognize individual performance	1.000
	When reporting, I consider student improvements	0.443
Induced competition	I encourage competition among my students	0.274
Eigenvalue: 1.267	I highlight the best performing students as an example	0.617
Cronbach's Alpha: 0.594	I explain to students how their performance compares to that of their peers	0.540
	I showcase the work of the students with the highest performance	0.628
	I give privileges to students who do the best work	0.352
Teaching efficacy	I can make students with difficulties understand me	0.005
Eigenvalue: 1.339	There are factors that influence my students more than I do	0.296
Cronbach's Alpha: 0.342	Some students will not make significant improvements	0.949
	I make a difference in the lives of my students	0.088
	I can deal with many of the learning disabilities	0.049
	I help students achieve improvements in their performance	-0.069
	There is little I can do to achieve performance improvements	0.580

4 Sample sizes

Table A.10 shows the number of students and groups in administrative and survey data at different moments of time. In 2015 the initial information from SAID is for a sample size of 38,006 students, CDIAR data contains information on 32,324 students and surveys were carried out throughout the 2015 academic year to 19,762 students.

Registration data refers to all students originally assigned to schools that are part of the experimental sample. The academic records sample is made up by the students who actually showed up at the beginning of the academic school year.

In 2015, a teacher survey was also carried out to math and language teachers in selected groups in all 171 schools in the experimental sample. This survey contains information about their type of contract, maximum level of education attained, teaching habits, among other information.

We also carried out a follow-up test at the end of the 2016-2017 academic year (end of grade 8) to the same subsample of students. The test had the same structure as the one used in seventh grade; all tracking students were surveyed, and some randomly-chosen students from bimodal and control were surveyed as well.

Appendix Table A.10: Data Sources by Cohort

	Administrative data			Survey data			
	Registration	Academic records		Endline		Follow up	
	Students	Students	Groups	Students	Groups	Students	Groups
2015 Cohort	38,006	32,324	907	19,762	657	16,778	652
Control	12,590	10,723	305	5,245	179	4,238	168
Tracking	12,695	10,812	300	9,288	300	7,716	293
Bimodal	12,721	10,789	302	5,229	178	4,824	191

Note: The sample of schools is 171 in 2015.

B Experiment

1 Compliance

Table B.1 shows the average percent of students who complied with treatment assignment, per treatment arm (Columns (1)-(3)). The mean in control and bimodal is close, and while the mean of the tracking groups is lower, column (4) shows that the difference between the three averages is not statistically significant.

Appendix Table B.1: Compliance

Time	Mean control (1)	Mean tracking (2)	Mean bimodal (3)	pvalues (4)	Observations (5)
7th Grade	0.921 [0.126]	0.888 [0.217]	0.923 [0.143]	0.162	907

2 Attrition

Between seventh and eighth grades 4,369 students left the schools in the experimental sample (these make up around 13.85% of students who were in the academic records during the 1st bimester of seventh grade). Column (1) of Table B.2 shows that this attrition is not correlated with being in tracking or in a bimodal school. Attrition for 9th grade is around the same, 8,025 are not in ninth grade (out of which 4,254 students had already left in eighth grade). Column (2) of table B.2 shows that this attrition is also not correlated with either treatment.

Appendix Table B.2: Student Attrition

Treatment	Attrited in Grade 8	Attrited in Grade 9
	(1)	(2)
Tracking	0.008	-0.010
	[0.007]	[0.010]
Bimodal	0.007	0.001
	[0.007]	[0.011]
Observations	32418	32418

Significance levels (* 10%, ** 5%, *** 1%) estimated using an OLS estimator controlling for strata fixed effects and allowing for clustered (school) correlation of the error term.

3 Group Stability

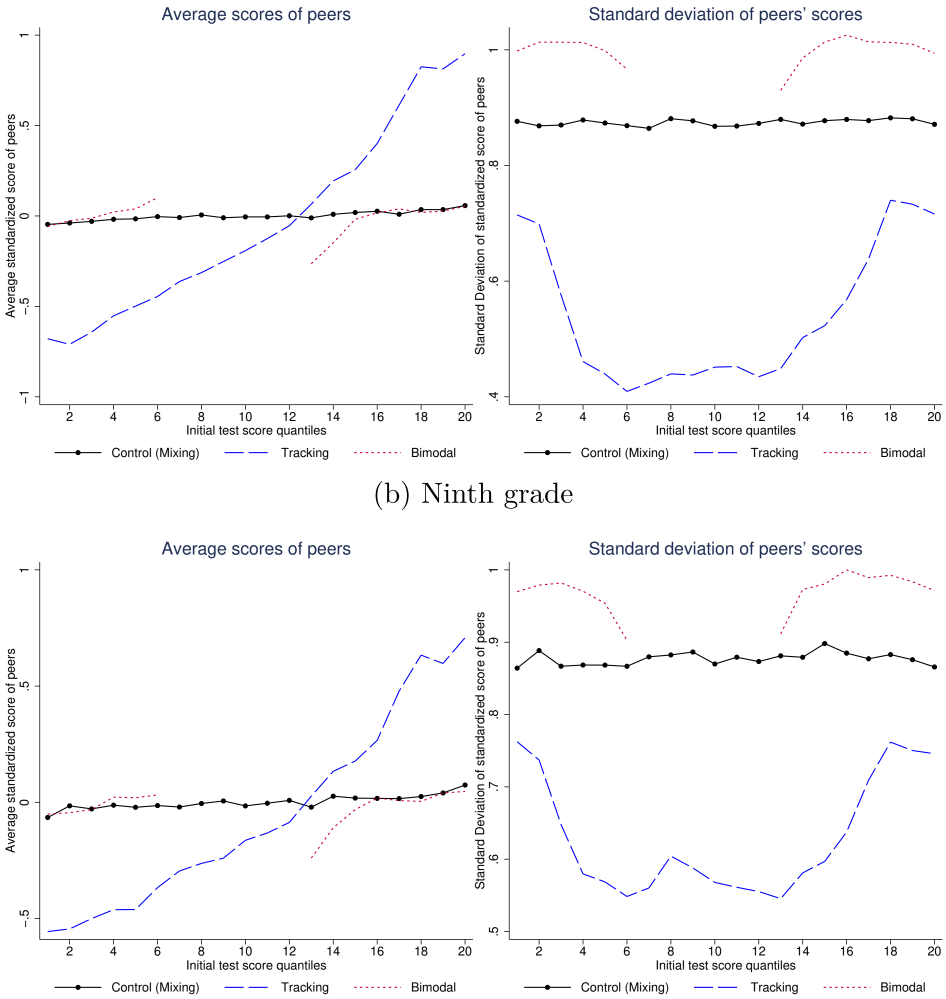
Table B.3 shows the average percentage of compliers in eighth and ninth grades in classrooms with respect to the initial number of students in those classrooms in seventh grade. The average percent of compliers in eighth grade was 63% and fell to 47% in ninth grade. These changes are similar between treatment arms and are not statistically different from one another (column (4)).

Appendix Table B.3: Group Stability

Time	Mean control	Mean tracking	Mean bimodal	pvalues	Observations
	(1)	(2)	(3)	(4)	(5)
8th Grade	0.646	0.622	0.636	0.750	907
	[0.280]	[0.305]	[0.292]		
9th Grade	0.468	0.449	0.479	0.659	907
	[0.281]	[0.295]	[0.290]		

Figure B.1 shows the changes in the average and standard deviations of peer's scores for different initial test score quantiles per treatment arm. We interpret this results as evidence of sustained exposure to tracking and bimodal throughout the whole middle school.

Appendix Figure B.1: Distributional changes for 8th and 9th grades



(a) Eighth grade

(b) Ninth grade